Quarterly Reviews of Biophysics 27, 4 (1994), pp. 435–481
Printed in Great Britain

Fundamentals of drug design from a biophysical viewpoint

WILFRED F. VAN GUNSTEREN, PAUL M. KING†, ALAN E. MARK

Laboratory of Physical Chemistry, Swiss Federal Institute of Technology Zürich, ETH Zentrum, CH-8092 Zurich, Switzerland
†Current address: Department of Chemistry, Birkbeck College, University of London, Gordon House, 29 Gordon Square, London WC1H 0PP, England

1. INTRODUCTION

Drug design means many things to many people. Commercially the aim is the development of compounds that can be patented and meet a variety of regulatory standards. In drug design, for medical purposes, toxicity and bio-availability are major considerations. Synthetically, questions related to ease of synthesis and chemical stability may dominate. From a physical perspective drug design is seen primarily as the process of optimizing specific (bio)molecular interactions. The biological activity of any compound can essentially be considered as a series of independent binding, transport and processing events. These events begin when the compound enters the body and end when it is either metabolized or excreted. Thermodynamically and kinetically each of these events may be expressed in terms of changes in free energy. The difference in free energy will determine how the compound partitions between different environments or between reactants and products and the intervening free energy barrier will determine the rate of such partitioning. Therefore, from a theoretical biophysical viewpoint drug design is concerned primarily with the estimation of the change in free energy for compounds in different environments. Most frequently it will be the difference between the free energy of a compound in water compared to that of the same compound bound to a specific protein receptor site that is of interest. Alternatively, it might be the difference in free energy of a compound bound to a bacterial or a mammalian form of the same enzyme. It may, however, correspond to the difference in free energy of a compound in an oxidizing or reducing intracellular environment where no specific macromolecular receptor can be identified. An example of such a case may be anti-cancer drugs designed to accumulate preferentially in rapidly metabolizing cells.

Free energy is a global property of a system and it is this that gives rise to the essential computational difficulty in drug design. Quantum mechanically, the Helmholtz free energy, F, of a system of N particles in a volume V at a temperature T, in terms of the canonical partition function, Q, is given by,

$$F(N, V, T) = -k_B T \ln Q(N, V, T)$$

$$= -k_B T \ln \left[\sum_j \mathrm{e}^{-E_j(N,V)/k_B T}\right] \qquad (1)$$

where the energy of a quantum mechanical state, j, of the system is given by $E_j(N,V)$, and k_B is Boltzmann's constant. The total free energy of a system is thus dependent on all possible electronic and nuclear degrees of freedom.

For any realistic system the absolute free energy cannot be calculated. It is only ever possible to estimate the change in free energy between two systems and then only by evoking a large number of often very crude assumptions or empirical

Abbreviations. CNDO, Complete Neglect of Differential Overlap; COMFA, Comparative Molecular Field Analysis; DNA, deoxyribose nucleic acid; HOMO, Highest Occupied Molecular Orbital; LFER, Linear Free Energy Relationship; LUMO, Lowest Unoccupied Molecular Orbital; MC, Monte Carlo; MD, molecular dynamics; MEF, Molecular Electric Field; MEP, Molecular Electrostatic Potential; QSAR, Quantitative Structure Activity Relationship; SAR, Structure Activity Relationship.

approximations. The relevant question for any given drug design problem is, for which sets of assumptions or approximations will the estimate of the change in free energy be useful? The answer will depend on the nature of the system, the amount of structural information available, the relative importance of quantum mechanical effects, the degrees of freedom in the system that can be safely ignored and the nature of the biological data against which the results can be compared.

This review aims to provide a broad overview of the range of methodologies that can be used to estimate relative free energies in drug design. The focus is not on applications, but on the physical bases of the methodologies, the assumptions on which they are based and the conditions under which such assumptions are valid. The primary purpose is to enable the reader to assess the applicability of a given class of methods to tackle a specific problem. Specific methods are only discussed by way of example. We do not aim to be encyclopaedic. Despite the fact that empirical methods currently dominate drug design the review is also intentionally biased toward the use of explicit free energy calculations based on detailed structural information. This bias merely reflects the fact that as the cost of computational resources falls the trend toward potentially more accurate explicit methods will be inevitable.

2. DRUG DESIGN BASED ON LEAD COMPOUNDS

All structure based drug design methods aim to derive a relationship between the topology or properties of a given molecular structure and a specific biological activity. Where the precise target or action of a lead compound is not known, such *structure activity relationships* (SAR's) must be derived empirically. The aim is to obtain a set of mathematical relationships between the properties of a set of related structures which form a given training set and some measure of biological activity that can later be used in a *quantitative* fashion to predict the activity of novel compounds (QSAR). In principle any physico-chemical or derived property that varies systematically in the series of test compounds can form the basis of a QSAR study.

The literature relating to the use of empirical methods in specific drug design studies is immense. A number of excellent general reviews exist (Martin, 1978; Franke, 1984; Martin, 1991; Silverman, 1992; Kubinyi, 1993) and it is not our intention to repeat the exercise here. Instead we wish to address the question of why certain molecular properties frequently show simple correlations to a variety of biological activities and under what conditions the assumptions that are evoked to account for such correlations might hold.

2.1 *Classical quantitative structure activity relationship (QSAR) approaches*

A century ago it was recognized that the narcotic effect of a series of neutral compounds appeared to be a function of their oil:water partition coefficient, P. Subsequently, the biological activity within many series of related compounds was shown to depend in a simple manner on the free energy of transfer from water to

a variety of organic phases given by the logarithm of the partition coefficient, log P. The basic assumption inspired by such studies was that the organic phase mimicked the interaction of the compounds with their site of action in biological membranes (Martin, 1978).

Modern QSAR studies stem, however, from work on the effects of sterically remote substituents on the rates of organic reactions which led Hammett (Hammett, 1940) to propose SAR's based on an empirical electronic parameter, σ, defined as

$$\log \frac{k_X}{k_0} = \rho \sigma_X \tag{2}$$

where k_0 is the rate or equilibrium constant of a reference compound, k_x the constant for the compound containing the substituent X and ρ is a proportionality constant characteristic of the sensitivity of the reference compound to substitution at a specific site. Taft (1956) extended the work of Hammett to include an additional parameter, E_s, defined in an analogous manner to the electronic parameter σ, to account for steric effects such that

$$\log \frac{k_X}{k_0} = \rho \sigma_X + \delta E_s \tag{3}$$

where δ is again a system dependent scaling parameter analogous to ρ. The approach illustrated by the work of Hammett and Taft was generalized by Hansch and coworkers (Hansch *et al.*, 1963; Fujita *et al.*, 1964; Hansch & Fujita, 1964) who proposed a group hydrophobicity parameter, π, defined in an analogous manner to σ and E_s based on octanol:water partition coefficients and introduced the assumption that the effects of the steric, electronic and hydrophobic properties of a given molecule on biological activity were independent and additive. This led to the general principle of *linear free energy relationships* (LFER's) where the activity, A, or the inverse of the effective concentration, $1/C$, could be expressed as

$$\log A = f_h(x_h) + f_e(x_e) + f_s(x_s) + \text{constant} \tag{4}$$

in which the log of the activity (a free energy) is assumed to be a linear combination of independent functions describing the hydrophobic, h, electronic, e, and steric, s, properties of a given compound. The functions, $f(x)$, are frequently assumed to be linear but may in principle take any form. Also, it should be noted that although originally expressed in terms of π, σ and E_s a myriad of other physical and derived parameters have been used in different QSAR studies (Franke, 1984).

LFER's and QSAR studies have proven to be very useful. From a physical perspective, however, it is not clear why LFER's should hold or what can be inferred in regard to the particular system when they do hold. The simplest type of biological activity is the direct binding of a test compound to an isolated receptor. In this case log A is directly proportional to the free energy of binding. Free energy is a global property of a system and cannot formally be separated into

a sum of components or group contributions unless the interactions on which the separation is based are either uncorrelated or purely enthalpic (van Gunsteren *et al.* 1993). For the interaction of specific substituents with specific residues of a receptor protein this will most likely not be the case. In contrast, in whole body or cellular assays the measured activity may reflect a sequence of uncorrelated transport and recognition events. The value of log $P_{\text{octanol:water}}$ might then reflect the rate of transport through the membrane and the oxidation potential of the compound its reactivity within the cell. In this case the application of (2) may well be valid.

The difficulty in assigning group free energies for use in QSAR studies can be illustrated in relation to the group hydrophobicity parameters of Hansch. Hansch defined the hydrophobicity π_X of a given substituent X as

$$\pi_X = \log P_X - \log P_H \tag{5}$$

where P_X and P_H are the octanol:water partition coefficients of a compound containing the substituent X and a hydrogen atom respectively. Octanol is often chosen as a reference solvent as, due to the presence of hydroxyl groups, it is assumed to approximate the hydrogen bonding environment of a biological membrane or protein. In the series of substituted aromatic compounds initially investigated by Hansch and coworkers specific π values were, in the absence of strong electronic effects, shown to be constant and additive. Hydrophobicity is, nevertheless, an essentially entropic phenomenon (Tanford, 1973). In this respect the additivity of π values is surprising. Both phases are, however, liquid. The nature of the local solvent environment in which the substituent is inserted, and hence the associated change in entropy, will depend primarily on the nearest neighbour atoms. For a series of aromatic parent compounds the nature and spatial arrangement of the neighbouring atoms is essentially constant. As expected, different hydrophobicity parameters or the use of correction terms are required for the same substituents attached to aliphatic chains or in close proximity to groups that perturb the local solvent environment. Using a variety of approaches log P values can be empirically predicted with high accuracy (Suzuki & Kudo, 1990). It is to be expected, however, that the use of group hydrophobicity parameters is more appropriate in liquid-like rather than highly structured environments.

2.2 *Empirical methods incorporating spatial information*

The interaction of a given compound with a specific receptor site will depend not only on the physical properties of the isolated substituents but also on their spatial arrangement. Crude spatial indices such as the steric parameters of Verloop *et al.* (1976), can be used to describe the volume and shape of a given substituent but such measures do not explicitly incorporate conformational information. To include explicit spatial information assumptions in regard to the active conformation and the mutual alignment of the test compounds must be made. In the distance geometry approach of Crippen (Crippen & Havel, 1988) such model

dependencies are minimized by considering only distances between atoms or interaction sites. Sets of distance bounds describing the conformational flexibility of each of the test compounds are generated. A comparison of the test compounds is then made based on the assumption that there is a single active conformation or pharmacophore. Sets of interaction sites are defined and geometrically allowed binding modes for each of the test compounds are evaluated. A simple scoring function based on favourable and unfavourable interactions with these sites can then be used to correlate structural features with some measure of biological activity (Ghose & Crippen, 1985). Alternatively, the ensemble of test compound configurations weighted by some measure of biological activity can be used to extract a set of common distances between potential interaction sites in order to define a potential pharmacophore or pseudo binding site (Sheridan *et al.* 1986). Although defining such a consensus binding site may be helpful in proposing alternate test compounds, it does not necessarily bear any relationship to a physical binding site with which the compounds under investigation might interact. This is especially true if the biological data with which it is correlated has not been derived from binding studies using an isolated receptor. Distance geometry and other 3D-QSAR methods depend strongly on the assumption that chemically related structures bind in a similar conformation and in the same orientation to a given receptor. While for simple rigid molecules this generally may be a very good assumption, it is certainly not always the case, as illustrated by three closely related elastase inhibitors which not only show different binding modes, but interact with different subsites (Mattos *et al.* 1994).

Distances between interaction sites is only one of a number of measures of molecular shape or molecular similarity that have been used to incorporate conformational information into QSAR studies. Other indices include steric volume overlap, charge matching and atom pair matching (Hopfinger & Burke, 1990). Similarity indices based on quantum mechanical calculations are discussed in the next section. Similarity indices discriminate on the basis of molecular conformation. Thus, results from such comparisons will depend on the model used to generate the three-dimensional structure of the test compounds. Structures can be generated using knowledge based approaches such as the programs CONCORD or WIZARD (Rusinko III *et al.* 1988; Leach *et al.* 1990), or minimum energy configurations from quantum mechanical or empirical force field calculations can be used. Similarity indices have also been derived as trajectory averages from molecular dynamics simulations. As with distance geometry methods it is assumed that there is a single active conformation and test compounds are superimposed on a given reference compound before comparison. This inevitably leads to a dependence on the choice of reference compound and the superposition criteria.

In the *comparative molecular field analysis* (CoMFA) method of Cramer III *et al.* (1988) molecules in a given test series are again aligned on a chosen parent structure. The potential energy with respect to a given force field is then sampled in the space surrounding each molecule using a regularly spaced grid and correlated at each point with a measure of biological activity. The correlation is

performed using the partial least squares method developed by Wold and co-workers (Wold *et al.* 1984) with cross-validation to give some measure of the predictive ability of the potential energy at each grid point. The result can then be expressed as a 3-dimensional contour surface reflecting the relationship between the molecular field and a specific biological activity.

CoMFA analysis is in general performed using the non-bonded terms from classical molecular mechanics force fields. The implementation in the molecular modelling package SYBIL standardly uses a 6–12 van der Waals and a Coulomb potential energy function, the latter with a distance dependent dielectric. The potential energy is calculated with respect to a probe atom. Although Cramer *et al.* (1988) initially used parameters for an sp^3 hybridised carbon atom carrying a charge of +1, the choice of probe atom and charge is essentially arbitrary and may be varied in order to optimize the correlation to a given set of experimental data. The use of such a pseudo physical force field has, however, led to ambiguities in the manner in which results from such studies should be interpreted. Specifically, there are questions remaining about whether the potential energy surface generated in the analysis reflects the structure of a specific receptor and whether the use of such a potential limits the analysis to purely enthalpic effects. Interaction energies obtained from static modelling with classical molecular mechanics force fields can at best only indicate enthalpic contributions to binding. It has been argued, therefore, that a CoMFA study using a potential energy function expressed purely in terms of van der Waals and Coulomb interactions cannot be expected to correlate with entropically driven phenomena or atom type specific interactions such as hydrogen bonding without the inclusion of special terms. In practice, inclusion of hydrophobic potentials (Kellogg *et al.* 1991) and/or an explicit hydrogen bonding term using for example the GRID potential energy function (Goodford, 1985; Kim, 1991) does not necessarily improve the overall correlation (Folkers *et al.* 1993; Kim *et al.* 1993). As an empirical method CoMFA is not limited to the use molecular mechanics force fields (Waller & Marshall, 1993). Conversely, the use of such force fields to extract correlations should not be confused with calculation of interaction energies based on the structure of the ligand-receptor complex. This was illustrated in a recent study of Klebe & Abraham (1993). Using crystallographic data to align a series of endothiapepsin inhibitors they demonstrated a significantly better correlation to enthalpic changes on binding than to entropy or free energy changes using a van der Waals and Coulomb potential energy function. The alignment correctly represented the binding to the receptor. Since the enthalpic changes were large and as enthalpic changes can be both formally separated into atomic contributions and equated to interaction energies, this result is not surprising. Fitted to free energies, however, the correlations generated by CoMFA can neither be used to infer interaction energies nor be expected to reflect details of the actual receptor. In the same study Klebe and Abraham observed for inhibitors of thermolysin that alignment based on crystallographic data yielded substantially inferior correlations to free energy data than two alternative alignment procedures. Where the activity data relates not to receptor studies but to cell or organ assays any

inferences drawn in regard to a specific receptor site from the generated correlations are even less reliable. In summary, although the potential energy terms commonly used in CoMFA have been derived in relation to detailed 3-dimensional structural information they are used in such studies simply as parameters. In this way the combination of potential energy terms from different sources, truncation or scaling of specific interactions, and fitting to non-linear functions can all be justified. In doing so, however, the biophysical basis of the force field is largely lost.

2.3 *Empirical methods based on the calculation and comparison of molecular quantum mechanical and electrostatic properties*

Quantum mechanical methods permit the calculation of a large number of molecular properties that can be empirically related to the action of a drug molecule. Quantum mechanical calculations can indicate preferred conformations, the distribution of charge within a molecule, possible tautomeric states, and potentially reactive functional groups (Richards, 1983). Such calculations are nowadays fairly routine with semi-empirical and *ab initio* quantum mechanical packages (for example Frisch *et al.* 1992; Stewart, 1990) and are very often the first step in understanding the behaviour and action of a potential drug molecule. Due to computational limitations and the fact that the relevant environment in which the molecule acts or the receptor to which it binds is often unknown many approximations must be made. Most commonly the calculations are performed in vacuo or alternatively in a dielectric continuum whose permittivity reflects that of the proposed environment. Alternatively a *supermolecule* calculation can be performed in which a few atoms of the environment (solvent molecules, amino acid residues etc.) are placed to mimic, at least approximately, the perturbing effect of the surroundings. The accuracy of quantum mechanical calculations is also limited by the number of electrons in the molecule which itself places limitations on the size of basis set that can be used (Hehre *et al.* 1986). Given that the relationship between the drug molecule's behaviour in isolation and at a receptor is uncertain, the use of the highest levels of theory incorporating large basis sets and electron correlation is not usually warranted. In cases where a whole range of molecules is systematically being studied such extensive calculations would in any case be prohibitively expensive. Calculations based on the quantum mechanical or electrostatic properties of molecules tend to ignore the entropic component of drug action, that is, the underlying assumption is that enthalpic terms dominate the activity. Thus drug design efforts are frequently aimed at maximizing binding capacity, particularly through complementarity of properties for the ligand and host believed to be involved in drug action (Dean, 1987).

2.3.1 *Conformational energy*

One of the most useful quantities that can be derived from quantum calculations is the energy of the molecule and how this changes with conformation. This can

give insights into the low energy conformations which the molecule might adopt at the receptor or alternatively a range of conformations which may not have minimal energy but which may be active and stabilized at the receptor. Such calculations are usually performed by systematically changing all the dihedral angles within the molecule, optimizing each structure and calculating the energy for each conformation. In cases where there are too many such dihedral angles, portions of the molecule are independently optimized and held rigid while rotation about a few critical angles is performed. The resulting information, namely the conformational energy, is usually contoured as a function of two dihedral angles. Conformational *free* energies can also be determined, which involve the additional calculation of vibrational frequencies for use in determining the vibrational partition function. A reaction mechanism can be postulated and the structure and energetics along a proposed reaction coordinate can be determined, including identification of the transition state. This may be relevant to drug action in cases where an enzyme is believed to stabilize the transition state of a reaction and one is interested in designing a transition state mimic to block the reaction.

The energies of simple reactions in vacuum such as protonation and tautomerism are straightforward to determine using quantum mechanical calculations. Such calculations may indicate whether alternative structures or protonation states are energetically feasible. The reactivity of a molecule can also be investigated by visualizing the frontier orbital electron density (Fukui *et al.* 1952). Electron density in the HOMO (highest occupied molecular orbital) will, in principle, indicate the position of electrophilic attack while for nucleophilic attack the LUMO (lowest unoccupied molecular orbital) is of importance. In order to compare frontier orbital electron density on different molecules the concept of *superdelocalizability* has been introduced. Here the contribution to the electron density from each molecular orbital is weighted by the energy of the orbital.

2.3.2 *Electron density*

Given a lead compound one usually wishes to compare it with a potential drug molecule so that the efficacy of the latter may be estimated. Perhaps the most natural choice for a quantum mechanical property with which to compare molecules is the electron density, $\rho(\boldsymbol{r})$. This is the number of electrons per unit volume and can be written in terms of the total molecular wavefunction Ψ as follows:

$$\rho(\mathbf{r}) = N \int \ldots \int |\Psi(\boldsymbol{x}_1, \boldsymbol{x}_2, \ldots, \boldsymbol{x}_N)|^2 \mathrm{d}s_1 \, \mathrm{d}\boldsymbol{x}_2 \ldots \mathrm{d}\boldsymbol{x}_N. \tag{6}$$

Here N is the total number of electrons of the system and $\boldsymbol{x}_i = (s_i, \boldsymbol{r}_i)$ are the spin and space coordinates of electron i. Integration of $\rho(\boldsymbol{r})$ over the whole space yields the total number of electrons in the system. The electron density lies at the heart of the currently fashionable density functional approaches to quantum chemistry (Parr & Yang, 1989) which follow from the realization that $\rho(\boldsymbol{r})$ determines the ground state wavefunction of the system and thus all its electronic properties. The electron density can also be expected to represent the underlying nuclear

framework of the molecule since its overall electron density can essentially be considered as the sum of spherical atomic electron densities with slight deformations. Electron density is a useful concept with which to work since it can be directly determined by experimental methods such as X-ray diffraction. The electron density is straightforward to compute using both *ab initio* and semi-empirical quantum mechanical programs. There are, however, limitations on the size of molecule and the number of basis functions used in the calculation. Semi-empirical methods can be used for a few hundreds of atoms while the highest accuracy *ab initio* methods are restricted to an order of magnitude fewer atoms. The easiest, and crudest, way of indicating the charge distribution within a molecule is to calculate atom-centred partial charges. This is the net charge, expressed in number of electrons, residing on an atom within a molecule and can be calculated from quantum mechanical or more empirical methods. Because the partial charge on an atom is not a physical quantity, i.e. it cannot be measured experimentally and there is no associated quantum mechanical operator, many methods of calculating it exist and assignment of charge is somewhat arbitrary. However, the partial atomic charges can indicate (i) whether a given atom has increased or decreased charge density relative to the unbound state and (ii) the relative size of build-up or depletion of charge on different atoms within the molecule. Inspection of a point charge distribution can indicate possible positions of nucleophilic or electrophilic attack.

Visualization of electron density can enable qualitative comparisons between molecules to be made. Very often two-dimensional contour plots are used to represent electron density in various planes of the molecule. Alternatively isodensity surfaces can be drawn around the molecules. These are sets of points in three-dimensional space at which the electron density attains a certain pre-set value, and are usually represented as a dot-surface, a triangulated mesh or an interpolated smooth surface. This can be used as a descriptor of the shape or surface of the molecule and can serve as a property to compare a given series of molecules or to map out the shape of an unknown receptor. Conversely one can define a certain molecular surface, such as the van der Waals or solvent-accessible surface, and evaluate the electron density at points evenly scattered over the area. The values of the electron density at these points can then be colour-coded to ease visualization. One might also analyze the *difference density* using similar methods. This is the difference between the molecular electron density and the density obtained by superposition of unperturbed atomic densities. Isodensity surfaces of the difference density clearly show how the electron distribution changes on formation of the molecule and can thus highlight regions of enhanced or diminished electron density. These may well be indicative of sites for electrophilic or nucleophilic attack, and one might envisage that such regions will occur in similar locations for molecules which bind to the same receptor and react by the same mechanism.

While qualitatively the display of electron density for a series of molecules can be instructive, to quantify the similarity between two molecular charge distributions some form of simple index is required. Similarity indices can serve

both as a criterion for the superposition of the molecules to aid in the mapping of an unknown receptor and as a descriptor for QSAR studies. One such similarity index was introduced by Carbó (Carbó *et al.* 1980). Given two molecules A and B, with respective electron densities $\rho_A(r)$ and $\rho_B(r)$, then a measure of the difference between the two densities is given by

$$\epsilon_{AB} = \int [\rho_A(r) - \rho_B(r)]^2 \mathrm{d}r = \int \rho_A(r)^2 \, \mathrm{d}r + \int \rho_B(r)^2 \, \mathrm{d}r - 2 \int \rho_A(r) \, \rho_B(r) \, \mathrm{d}r. \tag{7}$$

If both molecules remain rigid then only the last term of (7) varies as the relative position of the molecules changes. Thus superimposing the molecules by minimizing the difference of the charge densities corresponds to determining a maximum for the integral $\int \rho_A(r) \, \rho_B(r) \, \mathrm{d}r$. Carbó proceeded to define a normalized measure of similarity, r_{AB}, given by

$$r_{AB} = \frac{\int \rho_A(r) \, \rho_B(r) \, \mathrm{d}r}{\left(\int \rho_A(r)^2 \, \mathrm{d}r \right)^{\frac{1}{2}} \left(\int \rho_B(r)^2 \, \mathrm{d}r \right)^{\frac{1}{2}}} \tag{8}$$

where r_{AB} lies in the interval [0,1]. Molecules with complete similarity, $\rho_A(r) = \rho_B(r)$, will have $r_{AB} = 1$ while complete dissimilarity will be indicated by $r_{AB} = 0$. This was originally implemented within a semi-empirical CNDO framework although an improved *ab initio* formulation has been presented (Bowen-Jenkins *et al.* 1985). The less accurate quantum mechanical methods tend to be less discriminating than the more accurate ones, while the latter tend to be too expensive for optimization of the similarity. This similarity measure is also very dependent on the manner in which the molecules are superimposed. In the presence of heavy atoms the similarity is dominated by their large electron density close to the nucleus and small misalignment of such atoms can give rise to unrealistically low similarity values. To overcome this problem use of only the valence electron density has been proposed and found to give results more in tune with chemical intuition.

The Carbó index of electron density similarity is not unique and while r_{AB} is sensitive to the shape of the electron density it is not sensitive to its magnitude. For example, when $\rho_A(r) = n\rho_B(r)$ then r_{AB} is still equal to unity indicating full similarity. An alternative index, albeit in a different context, has been proposed by Hodgkin (Hodgkin & Richards, 1987):

$$s_{AB} = \frac{2 \int \rho_A(r) \, \rho_B(r) \, \mathrm{d}r}{\int \rho_A(r)^2 \, \mathrm{d}r + \int \rho_B(r)^2 \, \mathrm{d}r} \tag{9}$$

s_{AB} compares the shape of the electron density and also its magnitude. Given the condition $\rho_A(r) = n\rho_B(r)$ then $s_{AB} = 2n/(n^2+1)$ and can thus distinguish between the molecules. Despite valence-only calculations both indices still tend to be more

sensitive to the positions of the nuclei than to the long-range valence electron density. An innovative attempt to overcome this limitation was made by using densities in *momentum* space, $\rho(\boldsymbol{p})$, rather than *position* space (Cooper & Allen, 1989). The alternative similarity index

$$S_{AB}(n) = \frac{2\int p^n \rho_A(\boldsymbol{p})\rho_B(\boldsymbol{p})\,\mathrm{d}\boldsymbol{p}}{\int p^n \rho_A(\boldsymbol{p})^2\,\mathrm{d}\boldsymbol{p} + \int p^n \rho_B(\boldsymbol{p})^2\,\mathrm{d}\boldsymbol{p}} \tag{10}$$

was introduced, in which $p = |\boldsymbol{p}|$. This measure of similarity has a number of advantages. Firstly, it is independent of the distance between the molecules in position-space and so many of the problems associated with the superposition of the two molecules will be avoided. Secondly, $\rho(\boldsymbol{p})$ is dominated by the valence electrons, which have low $\boldsymbol{p}$ values, whereas $\rho(\boldsymbol{r})$ is dominated by core-electrons and thus the position of the nuclei. Furthermore, use of $S_{AB}(n)$ for $n = -1,0,1,2$ can be used to measure similarity between different regions of the electron density.

2.3.3 *Electrostatic potential*

A quantity that has perhaps received more attention than the electron density in studies of molecules and in QSAR is the electrostatic potential, $V(\boldsymbol{r})$, defined as follows

$$V(\boldsymbol{r}) = \frac{1}{4\pi\epsilon_0}\left[\sum_{i=1}^{N}\frac{Z_i}{|\boldsymbol{r}-\boldsymbol{R}_i|} - \int\left(\frac{\rho(\boldsymbol{r}')}{|\boldsymbol{r}-\boldsymbol{r}'|}\right)\mathrm{d}\boldsymbol{r}'\right]. \tag{11}$$

The first term represents the contribution from the N atomic nuclei of charge $\{Z_i\}$ situated at $\{\boldsymbol{R}_i\}$, while the second term is the contribution from the electronic charge density. $V(\boldsymbol{r})$, often called the molecular electrostatic potential (MEP), represents the energy of interaction between the molecule and a proton situated at $\boldsymbol{r}$, or if a point-charge of value q is situated at $\boldsymbol{r}$ the energy of interaction will be given by $qV(\boldsymbol{r})$. It should be stressed that the electrostatic potential is usually calculated for an isolated molecule and so using $qV(\boldsymbol{r})$ to evaluate an energy of interaction is a first-order approximation, neglecting cooperative effects such as polarization and charge transfer. However the MEP can be used to give a very clear indication of the areas of three-dimensional space where groups of a given charge will be attracted or repelled. One can thus use the MEP to map out the surface of a pseudo receptor based on the assumption that the receptor will have a complementary electrostatic potential to that of the ligand. Furthermore, by comparing the MEP's for a number of ligands one can propose areas of the molecules associated with binding to a receptor or possible areas for electrophilic or nucleophilic attack. The electrostatic potential can be calculated in most *ab initio* or semi-empirical quantum chemistry programs, although once again there are limitations on the size of molecule that can be treated. Very often when working with a large number of molecules or with very large molecules, such as

proteins or DNA fragments, where a fully quantum mechanical treatment is not possible, a point charge approximation is made and the following expression used

$$V(\boldsymbol{r}) = \frac{1}{4\pi\epsilon_0} \sum_{i=1}^{N} \frac{q_i}{|\boldsymbol{r} - \boldsymbol{R}_i|} \tag{12}$$

where $\{q_i\}$ are the net atomic charges situated on the nuclei.

As with electron density one can introduce a similarity index with which to compare molecular electrostatic potentials (Hodgkin & Richards, 1987). This can be of the same form as expressions (8) or (9) with the electrostatic potential at position $\boldsymbol{r}$, $V(\boldsymbol{r})$, replacing the electron density $\rho(\boldsymbol{r})$. In their implementation Hodgkin and Richards used a point charge model and evaluated the integrals required for the calculation of the similarity index numerically on a grid. Generally reasonable results could be achieved with a grid which extends 1·0 nm beyond the molecule and has a mesh of 0·1 nm. To avoid singularities associated with the evaluation of the MEP at the nuclear sites the 'inside' of the molecule was excluded from the calculation. Details associated with the grid extent and fineness, the method used to optimize molecular geometry and obtain atomic charges have been addressed (Burt *et al.* 1990) as too has the introduction of flexible fitting for optimization of the similarity index (Burt & Richards, 1990). The use of a 2- or 3-Gaussian expansion for $1/r$ (Good *et al.* 1992*a*) allows the grid-based determination of the electrostatic potential to be replaced by analytic evaluation which makes the calculation two orders of magnitude faster. Furthermore more robust methods of optimization can be used so that the fitting of molecules is less likely to become stuck in local minima, and because there is no singularity at the nucleus there is no need to exclude the molecular volume from the similarity calculation. This method has been widely applied to the screening of results from 3D-database searches (Good *et al.* 1992*b*) and to the calculation of similarity matrices for use in QSAR studies (Good *et al.* 1993).

2.3.4 *Electric field*

A further molecular property of use in analysing and comparing structures is the molecular electric field (MEF), defined as follows

$$\mathscr{E}(\boldsymbol{r}) = -\nabla V(\boldsymbol{r}). \tag{13}$$

The electric field is thus the negative of the gradient of the electrostatic potential and as such is a vector quantity. The molecular electric field is usually considered in two ways. Firstly, the scalar product $\mathscr{E}(\boldsymbol{r})\cdot\boldsymbol{\mu}$, where $\boldsymbol{\mu}$ is a permanent dipole moment of a second molecule, gives the energy of interaction of the dipole with the field of the original molecule. Such interactions can be important in the binding of a ligand to a receptor. Alternatively one can study the field vector itself which indicates the force on a proton placed at $\boldsymbol{r}$. The MEF of a receptor may indicate the route by which a ligand is guided towards the binding pocket. Being a vector quantity the field is rather more difficult to visualize than the electrostatic potential and electron density, and hence it has probably been less used in drug design studies. However, it is possible to visualize the electric field, either by

displaying the vector orientations at points on a regular lattice or by plotting field lines which indicate the motion of a freely translating proton in the region of the molecule. The MEF has also been used in quantitative studies of molecular similarity (Hodgkin & Richards, 1987).

3. DRUG DESIGN BASED ON RECEPTOR STRUCTURE

If the three dimensional structure of a drug target receptor is known the design problem is substantially different from that based on lead compounds. In this case direct methods can be used to estimate differences in binding free energy for different compounds as opposed to the use of purely empirical correlations. The design process itself is, therefore, often referred to as being *de novo*. The primary concern in de novo drug design is computational efficiency. The increased accuracy of predictions based on higher level theory must be offset against a smaller range of compounds that can be investigated for the same cost. The choice of method is also governed by the quality of the structural data and which degrees of freedom in the system can be safely neglected.

In cases where the structure of a biological target is known, e.g. the crystal structure of an enzyme critical to viral reproduction, but few or no potential inhibitors exist, one is faced both with the problem of determining where an active molecule may bind and proposing potential agonists or antagonists. In such a case it is common to fix the conformation of the receptor. This neglects the dynamics of the receptor and implicitly assumes that entropic or enthalpic changes associated with the accommodation of the receptor to a specific ligand are negligible. Although this is a big approximation and may result in significant artifacts from a computational perspective, it is frequently the only option that yields a tractable starting point for further investigations.

3.1 *Site identification*

Before an agonist or antagonist can be proposed potential binding sites must be identified. Although these often can be inferred from mutation studies or by inspection of the structure for cavities within the molecule or clefts lying on the surface, in large systems containing many such sites it can be difficult to definitively identify by inspection for example the enzymically active site. Calculation and visualization of the electrostatic potential and electric field, as described previously in regard to empirical approaches, can be used to aid this process. As the host molecule consists very often of many thousands of atoms, one usually assigns point charges to atoms and calculates the various properties classically. The molecular electrostatic potential (MEP) can indicate regions where positive or negative charge will bind favourably. The MEP of the active site is expected to be complementary to that of ligands which bind – this is one criterion for identification of the site. The MEF can indicate the force on a charge and thus the vector field can be particularly instructive in showing how a charged ligand is directed towards the binding site, thus helping to identify the site.

One step beyond a simple electrostatic view of the host molecule is to incorporate other force-field terms to represent additional intermolecular forces. A widely used and conceptually appealing method is that of Goodford and co-workers called GRID (Goodford, 1985). The interaction of a *probe*, e.g. a functional group typically found on a ligand, such as methyl ($-CH_3$), amino ($-NH_2$), carbonyl ($=O$), and hydroxy ($-OH$), or a water molecule, is calculated with the fixed macromolecule. The energy of interaction of the probe is calculated at the points of a three dimensional grid superimposed over the macromolecule using an empirical potential energy function. The resulting energy values are contoured in three dimensions and displayed graphically along with the macromolecule. Contours at negative energy are assumed to indicate attractive regions for the probe and should occur in the binding pocket for probes found on known ligands. The most negative regions indicate the most favourable binding locations. The potential energy function used by GRID consists of three terms: (i) a van der Waals 6–12 function, which can be considered as defining the shape of the probe, (ii) an electrostatic interaction, which is a Coulombic potential containing a distance-dependent dielectric permittivity related to the environment of the interacting atoms, and (iii) a hydrogen bonding term which is angle dependent and allows for some mobility of the hydrogen atoms and lone-pairs (Boobbyer *et al.* 1989; Wade *et al.* 1993; Wade & Goodford, 1993). There are a number of simplifications present in this approach, which also occur in some other force field methods. Entropy is totally neglected, pairwise additivity of terms is assumed, no polarization or redistribution of charge is permitted and the macromolecule is treated as a static entity. The method is widely used because the information that results is simple to visualize and straightforward to interpret. The most negative contour regions show where most favourable enthalpic binding of a functional group will occur. Contours at less negative values indicate the amount of movement a probe may undergo and still have favourable interactions. The closeness of the contours can indicate what forces the probe would be under and in what direction these act. Together these factors often allow one to identify a binding site and further help in the design of potential ligands that will bind to the site.

3.2 *Shape based docking*

The DOCK algorithm (Kuntz *et al.* 1982; DesJarlais *et al.* 1986; Shoichet *et al.* 1992) attempts to generate geometrically feasible alignments of ligands within a receptor site of known structure based on a detailed matching of molecular surfaces. The volume of the binding site and that of potential ligands are first represented as a series of spheres which fit into their respective solvent accessible surfaces (Connolly, 1983). The ligands are treated either as completely rigid or as a series of rigid sub-fragments. Inter-sphere distances are then scanned for matches within a given tolerance and the structures superimposed based on matched distances. This generates a large number of potential alignments that are ranked in accordance to a given scoring function. The default scoring function has

varied in each implementation. In the most recent implementation a lattice is placed covering the receptor site and scores are accumulated for each lattice point that lies within a given cutoff distance of a ligand atom (Shoichet *et al.* 1992). Large negative scores are recorded for atoms which overlap with the lattice points. Different cutoff distances can be chosen for different atom types to account for close contacts of atoms involved in hydrogen bonding. Other scoring functions including a mixture of hard sphere and hydrogen bonding terms, a continuous scoring function based on inter-atom distances or interaction energies based on the AMBER force field have been used with mixed success.

Shape complementarity algorithms such as DOCK discriminate on a detailed matching of molecular surfaces. Using conformations of the receptor and ligand derived from crystallographic studies of the complex such algorithms routinely reproduce the experimental structure. However, such methods are sensitive to changes in the geometry of either the ligand or the receptor. Using DOCK to scan a data base for possible inhibitors of thymidylate synthase 3 of the 25 best scoring compounds inhibited the enzyme at sub millimolar concentrations. Crystallographic studies of one of the resulting complexes revealed that the bound inhibitor was shifted by 0·6–0·9 nm, binding to a different region of the active site (Shoichet *et al.* 1993). In a more recent study on non-peptide inhibitors of HIV-protease a proposed inhibitor bound 0·48 nm away and rotated by 79° from the predicted location in a different conformation than the docked structure (Rutenber *et al.* 1993). In that these and other investigations have led to novel inhibitors in different systems the method has been an undoubtable success. In regard to the prediction of the location of the binding site, the orientation of the ligand within a given site, or specific hydrogen bonding contacts, purposes for which the method was designed, the results are so far discouraging.

3.3 *Fragment build-up*

An alternative to molecular docking algorithms are fragment build-up procedures in which potential ligand molecules are constructed from simple precursors. Such schemes have the advantage that completely novel compounds can be suggested. However, as with docking algorithms such as DOCK, the accuracy of these methods is determined primarily by the form of the scoring function or force field calculation that must be used to discriminate between potential ligands and binding modes. Because the methods generate a very large number of alternatives, the discriminating function is in general crude and factors such as potential flexibility in the receptor site are not considered.

The types of compounds that will be suggested and the sophistication of the discriminating functions that can be used depend to a large degree on the implementation of the method. The build-up procedure GROW (Moon & Howe, 1991) constructs peptide models from a user selected starting position by piecing together amino acid fragments in conformations that will interact most favourably with surrounding atoms in the proposed receptor site. GROW operates by sequentially attempting to add all fragments from a preconstructed library to a given seed. To score each compound an interaction energy between the receptor

and the test compound is determined based on the AMBER molecular mechanics force field and a correction term for ligand desolvation. A manageable number of test compounds is maintained by only propagating the ten lowest energy conformations each round. The combination of assuming a rigid receptor, fixed ligand geometries and scoring on the basis of an interaction energy that includes a highly non-linear 6–12 van der Waals term means that, like DOCK, GROW discriminates primarily on a detailed matching of molecular surfaces. For this reason GROW, as expected, faithfully reproduces the sequence and configuration of peptides co-crystallized with a given receptor. The method is, however, extremely sensitive to the choice of starting position which must either be selected by the user or obtained using an algorithm such as DOCK. For this reason GROW is primarily applicable to the extension of a pre-existing ligand or in proposing alternate amino acid residues for a pre-existing ligand.

An example of an alternative implementation of a build-up procedure is the program LUDI (Böhm, 1992 *a*, *b*). LUDI aims to position small molecules into clefts or cavities in a protein structure such that protein hydrogen bond donors or acceptors and possible hydrophobic contacts are satisfied. This is done by first defining a series of interaction sites. In the later implementation a rule based procedure is preferred to define sites, but methods based on hydrogen bond and hydrophobic distributions extracted from structural databases or a probe atom procedure such as GRID are also described. Distances between interaction sites are then used to select and dock molecules from a rigid fragment library. To generate the final compounds bridging groups are used to join fragments. Again this procedure will correctly suggest a ligand that has been co-crystallized with a given receptor but is later deleted from the structure. The advantage of a LUDI type build-up procedure over DOCK or GROW is that the initial placing of unbridged fragments is performed independently, and a degree of tolerance in the positioning of the fragments is inherent in the method. Thus, the method is more tolerant, but less discriminating on the basis of surface matching. LUDI uses a rule based scoring function to discriminate between potential ligands based on hydrogen bond geometries, volume overlap and buried surface area. Such rule based functions are essentially empirical and necessarily hold only for the range of compounds against which they are fitted.

Buildup procedures illustrated by GROW and LUDI in general attempt to grow or join molecular fragments based on known structures. Recently, however, a number of atom based methods have also been proposed (Nishibata & Itai, 1991; Rotstein & Murcko, 1993; Pearlman & Murcko, 1993). The most novel of these methods is that of Pearlman and Murcko, which attempts to dynamically allocate atom types and bonding topologies during a molecular dynamics simulation of a number of initially unconnected atoms, the motions of which are restrained to the volume of a proposed binding site. The atoms interact with the receptor via the AMBER molecular mechanics force field and Monte Carlo type moves are used to change atom types and bonded interactions. Although conceptually interesting, the computational cost of this method probably cannot be justified in terms of its ability to suggest and discriminate between potential inhibitors. Atom based build-up procedures can also lead to chemically unreasonable structures.

A. Model building	– packing considerations – matching of hydrogen bond donors and acceptors – matching of opposite charges – electric field evaluation	RIGID
B. Energy calculation	– systematic search (SS) – heuristic search (HS) – energy minimization (EM) – Monte Carlo simulation (MC) – molecular dynamics (MD) – stochastic dynamics (SD)	FLEXIBLE
C. Free energy calculation	– solvent effect – binding constant	FLEXIBLE ENTROPY

Fig. 1. Different levels of sophistication of molecular modelling.

4. DRUG DESIGN BASED ON RECEPTOR-LIGAND INTERACTIONS

When both the spatial structure of receptor and ligand and their relative position and orientation are approximately known, one may attempt to calculate a binding constant based on an evaluation of the atomic interactions between receptor and ligand atoms, both in the bound and unbound states. In such a calculation a variety of choices is to be made. Which (atomic) interaction sites are included in the (free) energy calculation? Which degrees of freedom (atomic, electronic) are explicitly used as variables in the calculation. For example, is a rigid model or a flexible model used? To which extent is the environment (solvent, ions, membrane) taken into account in the calculation, and in which manner, e.g. as a mean interaction or explicitly? Is only energy calculated or are entropic contributions to the free energy also included?

Different levels of sophistication with respect to the modelling of receptor-ligand interactions are distinguished in Fig. 1. The simplest level is indicated by the term model building, and involves an evaluation of the binding capacity of receptor and ligand based on spatial packing considerations, the matching of hydrogen bond donors and acceptors, of opposite electric charges or on an electric field evaluation, all using a rigid receptor-ligand structural model. For example, the electrostatic component of a binding energy may be estimated by calculating the electric field at the binding site due to charges on the receptor and then evaluating the energy of the ligand charges in this field. In such a calculation the precise conformations of receptor and ligand are not expected to play a significant role, since the electrostatic interaction is a slowly varying function of the distance between charges. It is the absence or presence of a charge which influences the field and energy, not its precise location at a distance from the binding site. However, when considering interactions such as the van der Waals, covalent bond-length and bond-angle interactions, which possess a much larger sensitivity to the distance between interaction sites, the interaction energy obtained will be sensitive to small changes in receptor and ligand structure. In this case the relative

energies obtained for different rigid receptor-ligand structures are not likely to correspond to the relative binding energies of the relaxed structures of the complexes.

The next level of sophistication of molecular modelling (Fig. 1) involves the incorporation of molecular flexibility as illustrated in Fig. 2. Flexibility can be taken into account by using an energy function or interaction potential $U_{phys}(\boldsymbol{r})$, which contains general physical information on molecular structure and flexibility. For example (van Gunsteren & Berendsen, 1990), an interaction function for biomolecular systems is

$$\begin{aligned} U_{phys}(\boldsymbol{r};s) = & \sum_{bonds\, n} \tfrac{1}{2}k_n^b [b_n - b_n^0]^2 + \sum_{angles\, n} \tfrac{1}{2}k_n^{ba} [\theta_n - \theta_n^0]^2 \\ & + \sum_{\substack{improper \\ dihedrals\, n}} \tfrac{1}{2}k_n^{id} [\xi_n - \xi_n^0]^2 \\ & + \sum_{dihedrals\, n} k_n^{da} [1 + \cos(m_n \phi_n - \delta_n)] \\ & + \sum_{pairs\, (i,j)} [B_{ij}/r_{ij}^{12} - A_{ij}/r_{ij}^6 + q_i q_j/(4\pi\epsilon_0 \epsilon_r r_{ij})]. \end{aligned} \tag{14}$$

It describes the energy of a molecular system as a function of the atomic coordinates of the N atoms of the system, indicated generally by $\boldsymbol{r}$. Expression (14) uses internal coordinates such as bond lengths b_n, bond angles θ_n, improper dihedral angles ξ_n and (proper) dihedral or torsional angles ϕ_n . The last term in (14) representing the nonbonded interaction is expressed in terms of the distance $r_{ij} = [(\boldsymbol{r}_i - \boldsymbol{r}_j)\cdot(\boldsymbol{r}_i - \boldsymbol{r}_j)]^{\frac{1}{2}}$ between atoms i and j. In the numerical practice the internal coordinates b_n, θ_n, ξ_n and ϕ_n are also expressed in terms of the cartesian coordinates $(\boldsymbol{r}_1, \boldsymbol{r}_2, \ldots, \boldsymbol{r}_N) \equiv \boldsymbol{r}$ of the N atoms. The functional form and the parameters $(b_n^0, k_n^b, \theta_n^0, k_n^{ba}, \xi_n^0, k_n^{id}, m_n, \delta_n, k_n^{da}, B_{ij}, A_{ij}, q_i, \epsilon_r) \equiv s$ of $U_{phys}(\boldsymbol{r};s)$ contain the general physical information on biomolecular systems: ideal bond lengths b_n^0, the variation of which is controlled by the size of k_n^b, partial atomic charges q_i , van der Waals parameters A_{ij} and B_{ij} , etc. They are chosen such that the function $U_{phys}(\boldsymbol{r};s)$ represents, as well as possible, the energy of a particular type of molecular system as a function of molecular configuration $\boldsymbol{r}$.

Molecular modelling using an energy function such as (14) involves three basic choices.

1. Which degrees of freedom are explicitly treated as variables, e.g. $\boldsymbol{r}_1, \boldsymbol{r}_2, \ldots, \boldsymbol{r}_N$ in $U_{phys}(\boldsymbol{r};s)$, and which are implicitly taken into account through the use of an effective interaction or potential of mean force for the explicit degrees of freedom? Such an effective interaction should contain the mean effect of the degrees of freedom which are not explicitly treated as variables in the molecular model, but which are averaged or treated as parameters. For example, the nuclear coordinates are often treated as parameters in a quantum mechanical

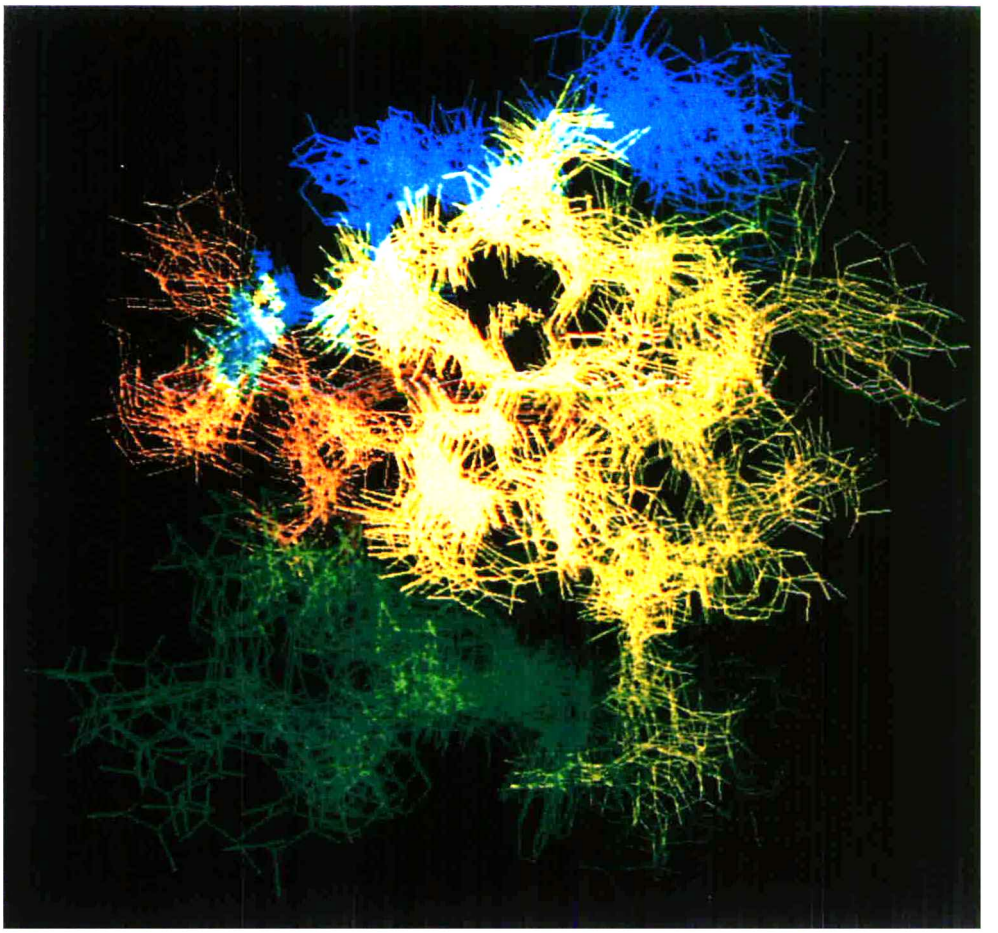


Fig. 2. An overlay of twenty configurations after a best fit superposition of all C^{α} atoms from a 100ps simulation of porcine insulin in aqueous solution illustrating the range of configurational space accessible to the protein on a subnanosecond time scale.

calculation of molecular properties using (part of) the electronic degrees of freedom as variables. Or, the parameters (s) of empirical classical interaction functions such as (14) are generally determined such that the interaction energy U_{phys} includes the mean effect of the (omitted) electronic degrees of freedom. A biomolecular system such as a receptor-ligand complex has in general more degrees of freedom (electronic, atomic) than can be reasonably treated as variables. Therefore, one has to select those degrees of freedom as variables in the molecular model, that are essential to a proper representation of the quantity or phenomenon one is interested in. With respect to the calculation of binding constants all degrees of freedom that are expected to give significant contributions to the energy or entropy of the complex compared to the individual receptor and ligand molecules, should be explicitly treated.

2. How is the interaction function for these degrees of freedom defined? This can be done at different levels of sophistication. Interaction functions of the

simplest type that are generally used in crystallographic modelling contain only terms describing covalent bond lengths, bond angles and chirality centers, and van der Waals repulsion between atoms. General empirical energy functions usually also contain terms representing dispersion energy and Coulomb energy (Gelin, 1993). A higher level of sophistication, without explicitly entering the realm of quantum mechanics, is reached by the inclusion of terms describing the electronic polarisability of atoms in the energy function. With respect to the calculation of binding constants such polarisation terms will be particularly important when a charged ligand is bound to a receptor. A mean treatment of electronic polarisation as in (14) may not be sufficient to obtain accurate binding constants in such a case.

3. How are the explicitly treated degrees of freedom sampled? Biomolecular complexes constitute microscopic systems, the properties of which are governed by the laws of statistical mechanics. This means that the probability of occurrence of a molecular (receptor-ligand) configuration $\boldsymbol{r}$ with energy $U_{phys}(\boldsymbol{r};s)$ is proportional to its Boltzmann factor

$$P(\boldsymbol{r}) \alpha \exp\left[-U_{phys}(\boldsymbol{r};s)/k_B T\right]. \tag{15}$$

A set of structures or configurations, in which each particular structure $\boldsymbol{r}$ occurs with relative probability $P(\boldsymbol{r})$ according to (15) is called a Boltzmann ensemble or distribution of configurations. The third basic choice of molecular modelling using energy functions involves the method by which a set of molecular configurations is generated (Fig. 1). For molecular complexes with a few explicit degrees of freedom these may be systematically varied (or searched:SS) to find the configurations with the lowest energy $U_{phys}(\boldsymbol{r};s)$, which will have the highest probability of occurrence $P(\boldsymbol{r})$. If the number of degrees of freedom grows, a systematic search of the vast configurational space of the molecular complex becomes impossible. Then, heuristic search (HS) methods, which visit a tiny, but hopefully representative set of configurations, are the only way to sample the Boltzmann ensemble. Though a simple method, energy minimisation (EM) is a very poor searching and sampling method, since it produces only one configuration which is a local minimum close to the initial structure. Much more powerful are methods such as Monte Carlo (MC) (Frenkel, 1993), molecular dynamics (MD) or stochastic dynamics (SD) (van Gunsteren, 1993) simulation, or combinations of these, that directly generate a Boltzmann ensemble. Since each configuration $\boldsymbol{r}$ in such an ensemble occurs with probability (15), averages and higher moments of the distributions can directly be calculated over the ensemble using a weight factor 1. A variety of even more powerful searching and sampling methods are known (van Schaik *et al.* 1992, 1993; Scheraga, 1993; Huber *et al.* 1994), which generally do not produce a Boltzmann ensemble of configurations. When taking averages of physical quantities over such a (non-Boltzmann) set of configurations, each configuration should be given the weight (15) in order to obtain a Boltzmann average. The calculation of binding constants should always be based on a Boltzmann average.

The next level of sophistication of molecular modelling involves the explicit evaluation of entropy or free energy (Fig. 1). From a Boltzmann ensemble the statistical equilibrium averages can be obtained for any desired property of the molecular system for which a value can be computed for each configuration of the ensemble. Examples of such properties are the potential or kinetic energy of (parts of) the system, structural properties, electric fields, etc. A number of thermodynamic properties can be derived from such averages. However, two important thermodynamic quantities, the entropy and the (Gibbs or Helmholtz) free energy, generally cannot be calculated using a statistical average. They are global properties of the molecular system that depend on the extent of configuration (or phase) space accessible to the system. Therefore, computation of the absolute free energy of a molecular system is virtually impossible. Yet, quantities important to drug design, such as binding constants and solubilities, are directly related to the free energy. Fortunately, over the past decade statistical mechanical procedures have evolved for evaluating *relative* free energy differences. They are rather demanding as far as computing power is concerned, but are very applicable in drug design based on receptor-ligand interactions.

4.1 *Structure and energy calculations using flexible 3D-structure-based models*

Molecular modelling of receptor-ligand complexes based on the matching of receptor and ligand properties or using rigid 3D-structures of receptor and ligand will produce a first estimate of the binding constant. Whether such an estimate bears any relation to reality will depend on the characteristics of the molecular system. For the binding of a rigid, positively charged receptor to a relatively rigid, negatively charged receptor (protein) it may be a reasonable estimate. However, for highly flexible molecules such as antibodies and antigens such an estimate is likely to be useless. In such cases flexibility of receptor, ligand and solvent has to be accounted for. Fig. 2 gives an impression of the flexibility of the protein insulin.

4.1.1 *Flexibility of ligand and receptor*

By using an energy function such as (14) for the atoms of the receptor and ligand, molecular flexibility can be accounted for in the drug design procedure. However, it depends on the technique used to search and sample molecular configurations, whether the potential molecular flexibility inherent to the energy function will be reflected in the energies and structures of the receptor-ligand complex that are obtained. The most commonly used search and sampling techniques are energy minimization, which only relaxes local strain in a molecular structure, and MC or MD simulation combined with high temperature annealing to extend the search (Goodsell & Olson, 1990; Kuntz, 1992; Stoddard & Koshland, 1992; Hodgkin *et al.* 1993).

Although the introduction of flexibility will in general improve the molecular model, it will only do so if the assumption of molecular rigidity is the accuracy limiting factor. If this is not the case the use of e.g. MC sampling techniques to sample conformations of flexible protein sidechains will not produce better

agreement with experimental data than simple modelling based on geometrical properties of amino acid sidechains (van Gunsteren & Mark, 1992*b*).

4.1.2 *Solvent effects*

An obvious way to limit the number of explicitly treated degrees of freedom in a biomolecular system is to omit all or almost all solvent degrees of freedom. Due to the abundance of solvent degrees of freedom for a biomolecule in solution, omission of these in an energy calculation easily reduces the required computing power by a factor of 10 to 50. Therefore, most drug design studies are carried out for molecules in *vacuo*. The complete neglect of solvent effects will limit the accuracy of the calculated properties, such as binding constants. So, what is the role of solvent molecules in a biomolecular system? The structure and stability of a molecular complex may depend on the type of solvent. Individual water molecules may play a structural role in e.g. protein-ligand or inhibitor binding. Polar solvents exert a dielectric screening effect on interactions between charges on the receptor and ligand, and the viscosity of the solvent will influence the dynamics of the atoms of the molecular complex, and may also influence the kinetics of the binding process.

Due to the many different low energy configurations occurring in a liquid, a solvent cannot be characterized by one or a few molecular configurations, but a (Boltzmann) ensemble of solvent configurations should be generated. So, when solvent degrees of freedom are included in the energy function (14), simulation techniques such as MC or MD are to be used, and systematic search (SS) and energy minimization (EM) are useless.

When simulating a microscopic system of finite size, the boundary of the system should be treated such as to minimize edge effects. The standard procedure is to use periodic boundary conditions. The solute or molecular complex and the surrounding solvent molecules are put into a periodic space-filling box, which is treated as if it is surrounded by identical translated images of itself. In this way basically an infinite periodic system is simulated. The periodic box should be taken large enough to avoid interactions between molecules and their periodic images, as illustrated in Fig. 3. This condition leads to sizeable amounts of solvent molecules, typically a few hundreds to thousands, to solvate the receptor, ligand or the complex, which makes such simulations expensive (van Gunsteren & Mark, 1992*a*).

An alternative to the explicit treatment of solvent molecules in a biomolecular simulation is an implicit one: the influence of the solvent on the solute degrees of freedom is incorporated in the energy function of the latter in an average manner. A potential of mean (solvent) force is used for the solute. Different mean solvation models are in use (van Gunsteren *et al.* 1994*b*).

1. *Accessible surface area type models*

   In this type of mean solvation model the local solvent contribution to the potential of mean force for solute atoms is taken to be proportional to the area of the solute atom or group of atoms that is accessible to solvent molecules (Ooi *et al.* 1987; Still *et al.* 1990; Schiffer *et al.* 1992; Wesson & Eisenberg, 1992).

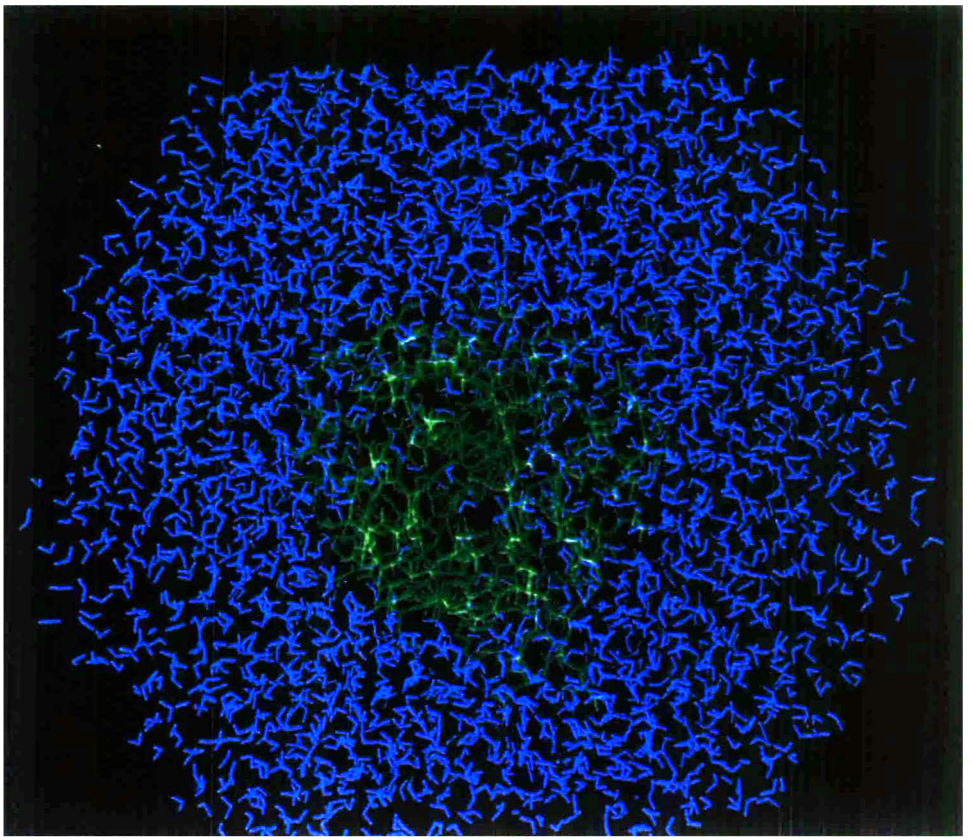


Fig. 3. The central or parent simulation cell from a simulation of hen egg white lysozyme (green) solvated by 5300 SPC water molecules with spherical (truncated octahedron) periodic boundary conditions illustrating the amount of solvent required to approximate infinite dilution in a periodic system.

Other local quantities, such as the hydration volume (Kang *et al.* 1988; Vila *et al.* 1991), or the number of solute-solvent contacts (Stouten *et al.* 1993) can be used too. The expression for the accessible surface area of a solute atom generally depends on the coordinates of the solute atom itself and those of its nearest neighbour solute atoms. Thus, a mean force potential based on an accessible area model becomes a many-body interaction which generally involves non-negligible computational effort. Currently, it is not clear whether this increased effort is offset by the limited accuracy of such mean solvation models in mimicking solvent effects.

2. *Simple pairwise solvation force models*

The computation of the mean force of solvation would be considerably simplified and sped up, if the mean force could be formulated as a sum of pairwise (two-body) interactions. In fact, the mean force potential in the solvent contact or occupancy model (Stouten *et al.* 1993) can be expressed as a sum of two-body terms of the form $exp(-r_{ij}^2/2\sigma^2)$, where r_{ij} is the distance between particles i and j. A slightly different functional form has been proposed

by van Gunsteren *et al.* (1994*b*). Although the use of a simple pairwise solvation force induces only a minor increase of the required computing effort, it remains to be investigated whether its accuracy is comparable to that of a simulation including a (thick) layer of solvent molecules.

3. *Dielectric screening models*
   Different approximate models for treating the long-range solute-solvent interactions due to dielectric screening and polarization effects are available too (Still *et al.* 1990; Solmajer & Mehler, 1991; Sharp, 1991, 1993; Gilson & Honig, 1991). Still *et al.* (1990) use an expression involving only one-body and two-body terms, which is based on the continuum approximation of a dielectric medium. A simple approach is to make the relative dielectric permittivity, ϵ, distance dependent, for example using a linear (Pickersgill, 1988) or sigmoidal function (Solmajer & Mehler, 1991) of the atom-atom distance. A more complicated way to incorporate long-range electrostatic effects using a continuum representation of the solvent is based on solving the Poisson-Boltzmann equation on a 3-dimensional grid (Sharp, 1991), from which a simple two-body interaction term is derived that accounts for charge-solvent interactions in an average way (Gilson & Honig, 1991). For reviews on the treatment of long-range electrostatic interactions in molecular simulations we refer to (Sharp, 1993; Berendsen, 1993; Smith & van Gunsteren, 1993).

Generally, the structure and stability of a molecular complex in aqueous solution will depend on the ionic strength of the solution. This implies that the presence of ions in the solvent should either be explicitly simulated (de Vlieg *et al.* 1989) or accounted for in an average manner by e.g. a Poisson-Boltzmann continuum model (Smith & van Gunsteren, 1993). Although the long-ranged interactions between ions in aqueous solution can be adequately approximated in a simulation, the explicit inclusion of ions in a simulation of a receptor-ligand complex in aqueous solution may not yield reliable results, since the relaxation time of a ionic distribution is likely to be longer than the simulation period. This slow relaxation is caused by the slow diffusion of hydrated ions in solution. In a biomolecular simulation including water and counterions, the simulation averages may then be easily based on non-equilibrated ion distributions, causing sizeable deviations from the mean effect of the ions. Therefore, the mean influence of the ionic solution might be better approximated by a simulation which only includes solvating water molecules.

4.1.3 *Incorporation of experimental data in a simulation*

If the molecular model and energy function $U_{phys}(\mathbf{r};s)$ are perfect and if a simulation can be carried on for an infinitely long time, the generated ensemble of molecular configurations will exactly represent the real molecular system: the structural and energetic properties derived from such an ensemble will be correct. In practice, neither condition is fulfilled. Atomic interaction functions, such as (14), are not infinitely accurate due to various approximations with respect to electronic degrees of freedom, quantum mechanical effects, many-body interactions, etc. that are made in their derivation. Second, an MD or MC

computer simulation cannot be carried out for infinitely many steps; typically for about 10^6 steps, which is in many cases not enough to sample all relevant molecular configurations of low energy.

One way to improve the reliability of a molecular simulation is to incorporate in the simulation experimental information on the particular molecular system that is simulated (van Gunsteren *et al.* 1994 *a*). This can be done by adding to the standard physical energy function $U_{phys}(\boldsymbol{r};s)$ an extra term $U_{restr}(\boldsymbol{r};s)$ which restrains or influences the motion of the system such that the generated trajectory or ensemble yields average properties in accordance with the experimental information on the specific molecular system. Thus, the energy function of the molecular system becomes

$$U(\boldsymbol{r};s) = U_{phys}(\boldsymbol{r};s) + U_{restr}(\boldsymbol{r};s). \tag{16}$$

The form of the restraining potential or penalty function $U_{restr}(\boldsymbol{r};s)$ depends on the type of experimental information, and should be chosen such that its value increases the more the property, calculated as an average over the trajectory or ensemble, deviates from the experimental value. Generally, a simple quadratic function of the difference between the simulated average and the measured value of a property is used for $U_{restr}(\boldsymbol{r};s)$.

For example, a penalty function which restrains the amplitudes of the averaged structure factors, $|<F_{calc}(hkl)>|$, to the observed ones, $|F_{obs}(hkl)|$, from an X-ray diffraction experiment is (Gros *et al.* 1990)

$$U_{restr}(\boldsymbol{r};k^{sfr},k_{sc},F_{obs}) = \tfrac{1}{2}k^{sfr} \sum_{reflections\,hkl} [|F_{obs}(hkl)| - k_{sc}|\langle F_{calc}(hkl)\rangle|]^2. \tag{17}$$

The summation runs over the reciprocal lattice vectors, the symbol $\langle \ldots \rangle$ denotes a trajectory or ensemble average, k_{sc} is a scaling factor to match the units of F_{calc} and F_{obs} and k^{sfr} in the force constant that determines the weight of the term U_{restr} with respect to the term U_{phys} in (16).

Other examples of restraining potential energy functions U_{restr} are found in the literature concerning structure determination based on NMR data. Nuclear Overhauser Effect (NOE) intensities can be converted to a set of upper bounds $\{r_{ij}^{ub}\}$ to the distances between atoms (hydrogens) i and j, which leads to a so-called distance restraining function (Kaptein *et al.* 1985).

$$U_{restr}(\boldsymbol{r};k^{dr},r^{ub}) = \tfrac{1}{2}k^{dr} \sum_{NOE\,pairs\,(i,j)} [\mathrm{MAX}(0,\langle r_{ij}^{-3}\rangle^{-1/3} - r_{ij}^{ub})]^2, \tag{18}$$

where the function MAX delivers the largest of its two arguments. Similarly, $\mathcal{J}$-coupling constant values can be restrained to the observed ones $\mathcal{J}^{obs}$ by the restraining function (Torda *el al.* 1993)

$$U_{restr}(\boldsymbol{r};k^{Jr},\mathcal{J}^{obs}) = \tfrac{1}{2}k^{Jr} \sum_{\substack{torsional\\ angles\,\phi_i}} [\langle \mathcal{J}(\phi_i(\boldsymbol{r}))\rangle - \mathcal{J}_i^{obs}]^2. \tag{19}$$

The $\mathcal{J}$-coupling constant is dependent on the torsional angle involving the three

covalent bonds connecting the atoms for which the $\mathcal{J}$-value is measured. Also chemical shifts σ can be restrained to the observed ones σ^{obs}, e.g. by the restraining function (Harvey & van Gunsteren, 1993)

$$U_{restr}(\boldsymbol{r}; k^{\sigma r}, \sigma^{obs}) = \tfrac{1}{2} k^{\sigma r} \sum_{resonances\, i} [\langle \sigma_i(\boldsymbol{r}) \rangle - \sigma_i^{obs}]^2 \tag{20}$$

where the summation runs over the resonances i. An example of the incorporation of microwave spectroscopic information is found in (King *et al.* 1993).

The average $\langle \ldots \rangle$ in (17–20) can be taken as a time (trajectory)-average (Torda *et al.* 1989) or as an average over space, i.e. different molecules (Scheek *et al.* 1991). In MD or SD simulations the use in (17–20) of the time average

$$\langle Q \rangle = \bar{Q}(t) \equiv t^{-1} \int_0^t Q(\boldsymbol{r}(t'))\, \mathrm{d}t' \tag{21}$$

of a quantity $Q(\boldsymbol{r}(t'))$ that depends on the molecular coordinates $\boldsymbol{r}$, is the natural choice. Formula (21) is the true average of Q and is used in the analysis of MD or SD trajectories, but it is not suitable for deriving a (restraining) force during the finite time of a simulation. As time increases, and $\bar{Q}(\boldsymbol{r}(t))$ is calculated over a longer period, the average (21) becomes less sensitive to instantaneous fluctuations. This problem can be avoided by building a decay into the summation over time with a characteristic decay time, τ, so that

$$\langle Q \rangle = \bar{Q}(t; \tau) \equiv [\tau(1 - \exp(-t/\tau))]^{-1} \int_0^t \exp(-t'/\tau)\, Q(\boldsymbol{r}(t - t'))\, \mathrm{d}t' \tag{22}$$

is used in (17–20) for the ensemble average in U_{restr}.

The ensemble average in (17–20) could also be implemented as an average over space, that is, over M different molecules

$$\langle Q \rangle = \langle Q \rangle_M \equiv \sum_{m=1}^{M} \mathrm{e}^{-E(\boldsymbol{r}_m)/k_B T}\, Q(\boldsymbol{r}_m) \Big/ \sum_{m=1}^{M} \mathrm{e}^{-E(\boldsymbol{r}_m)/k_B T}, \tag{23}$$

where the summation runs over equivalent molecules or systems which have a configuration denoted by $\boldsymbol{r}_m$. This expression assumes that the configurations of the M molecules are uniformly distributed. The advantage of time averaging over space averaging is that the relative Boltzmann probability of the configurations of a trajectory is guaranteed when proper equations of motion are integrated.

Finally, we note that the procedure to incorporate experimental information in an average manner, as sketched here, can be applied to other than the molecular properties discussed here. It could be used to force a receptor-ligand complex to adopt on average a given set of properties.

4.2 *Free energy calculations using flexible, 3D-structure-based models*

In terms of the energy, E, and the entropy, S, the Helmholtz free energy, F, of a system of N particles in a volume V at a temperature T is given by

$$F(N, V, T) = E(N, V, T) - TS(N, V, T). \tag{24}$$

In terms of the canonical partition function, $Z(N,V,T)$, it is given by

$$F(N,V,T) = -k_B T \ln Z(N,V,T)$$
$$= -k_B T \ln\left[(h^{3N} N!)^{-1} \iint e^{-H(\boldsymbol{p},\boldsymbol{r})/k_B T}\, d\boldsymbol{p}\, d\boldsymbol{r}\right] \tag{25}$$

where k_B is Boltzmann's constant, h is Planck's constant and the Hamiltonian,

$$H(\boldsymbol{p},\boldsymbol{r}) = \sum_{i=1}^{N} \boldsymbol{p}_i^2/(2m_i) + U(\boldsymbol{r}), \tag{26}$$

expresses the total energy of the system in terms of the coordinates $\boldsymbol{r} = (\boldsymbol{r}_1,\boldsymbol{r}_2,\ldots,\boldsymbol{r}_N)$, and the conjugate momenta, $\boldsymbol{p} = (\boldsymbol{p}_1,\boldsymbol{r}_2,\ldots,\boldsymbol{p}_N)$ of the N particles of the system. The masses of the particles are indicated by m_i and their interaction function by $U(\boldsymbol{r})$. From expression (25) it can be readily appreciated that the free energy is a global property given by a double integral of the (positive) Boltzmann factor $exp[-H(\boldsymbol{p},\boldsymbol{r})/k_B T]$ over all possible values of $\boldsymbol{p}$ and $\boldsymbol{r}$ which define the volume of phase space accessible to the system. Since MD or MC simulations of large molecular systems necessarily sample only a limited set of configurations, calculation of the free energy via (25) using a set of MD or MC configurations will suffer in a systematic way (F being too large) from incomplete sampling of phase space. Each additional part of phase space that is included in the integral will give a negative contribution to the free energy and a positive contribution to the entropy, since the natural logarithm is a monotonically increasing function of its argument. Even if the simulated configurations are representative of the complete ensemble and thus reasonably accurate ensemble averages may be calculated, the integral in (25) will not be accurate.

Re-expressing the free energy as a function of an ensemble average shows the problem of the practical calculation of free energy in a different manner. Integrating over the momenta $\boldsymbol{p}$ in the partition function and using $\int d\boldsymbol{r} = V^N$, we find

$$F(N,V,T) = -k_B T \ln\left[\frac{\int e^{-U(\boldsymbol{r})/k_B T}\, d\boldsymbol{r}\, V^N}{\int e^{+U(\boldsymbol{r})/k_B T} e^{-U(\boldsymbol{r})/k_B T}\, d\boldsymbol{r}}\right] - k_B T \ln\{[2\pi m k_B T/h^2]^{3N/2}/N!\}$$
$$= +k_B T \ln\langle e^{+U/k_B T}\rangle_{N,V,T} - k_B T \ln\{V^N [2\pi m k_B T/h^2]^{3N/2}/N!\} \tag{27}$$

where the ensemble average $\langle Q\rangle$ of a microscopically defined quantity $Q(\boldsymbol{p},\boldsymbol{r})$ is defined by

$$\langle Q\rangle_{N,V,T} \equiv \frac{\iint Q(\boldsymbol{p},\boldsymbol{r})\, e^{-H(\boldsymbol{p},\boldsymbol{r})/k_B T}\, d\boldsymbol{p}\, d\boldsymbol{r}}{\iint e^{-H(\boldsymbol{p},\boldsymbol{r})/k_B T}\, d\boldsymbol{p}\, d\boldsymbol{r}}$$
$$= \iint Q(\boldsymbol{p},\boldsymbol{r})\, P(\boldsymbol{p},\boldsymbol{r})\, d\boldsymbol{p}\, d\boldsymbol{r}, \tag{28}$$

where $P(\boldsymbol{p},\boldsymbol{r})$ is the probability of finding the system in the state characterized by $\boldsymbol{p}$ and $\boldsymbol{r}$. In an analogous manner the entropy, S, and the energy, E, can be expressed in terms of ensemble averages as

$$S(N,V,T) = -\frac{\partial F}{\partial T}$$

$$= \langle H\rangle_{N,V,T}/T - k_B \ln \langle e^{+U/k_B T}\rangle_{N,V,T}$$

$$+ k_B \ln\{V^N[2\pi m k_B T/h^2]^{3N/2}/N!\} \tag{29}$$

and

$$E(N,V,T) = \langle H\rangle_{N,V,T}. \tag{30}$$

Accurate calculation of the free energy and entropy is not possible due to the occurrence of the ensemble average $\langle\exp[+U/k_BT]\rangle$ in (27) and (29). Since the probability $P(\boldsymbol{r})$ of a molecular configuration is proportional to the Boltzmann factor $\exp[-U(\boldsymbol{r})/k_BT]$, it is small when the function to be averaged, $\exp[+U(\boldsymbol{r})/k_BT]$, is large and vice versa. For other quantities, e.g. the energy E, this problem does not occur when evaluating the ensemble average in (30).

In view of these difficulties, the calculation of the absolute free energy or entropy of a molecular system is virtually impossible. However, over the last decade statistical mechanical procedures to evaluate relative free energies have evolved, which are applicable to molecular systems of biochemical interest. For reviews of these techniques we refer to (Beveridge & DiCapua, 1989; King, 1993; van Gunsteren *et al.* 1993; Straatsma *et al.* 1993).

4.2.1 *Free energy differences by thermodynamic integration*

Using the so-called coupling parameter approach the difference in free energy between two states A and B of a system can be determined, if the Hamiltonian of the system is made a function of a coupling parameter λ, thus $H(\boldsymbol{p},\boldsymbol{r};\lambda)$, such that when $\lambda = \lambda_A$ the system corresponds to state A, $H(\boldsymbol{p},\boldsymbol{r};\lambda_A) = H_A(\boldsymbol{p},\boldsymbol{r})$, and when $\lambda = \lambda_B$, the system corresponds to state B, $H(\boldsymbol{p},\boldsymbol{r};\lambda_B) = H_B(\boldsymbol{p},\boldsymbol{r})$. The partition function

$$Z(N,V,T;\lambda) = (h^{3N}N!)^{-1}\iint e^{-H(\boldsymbol{p},\boldsymbol{r};\lambda)/k_BT}\,d\boldsymbol{p}\,d\boldsymbol{r} \tag{31}$$

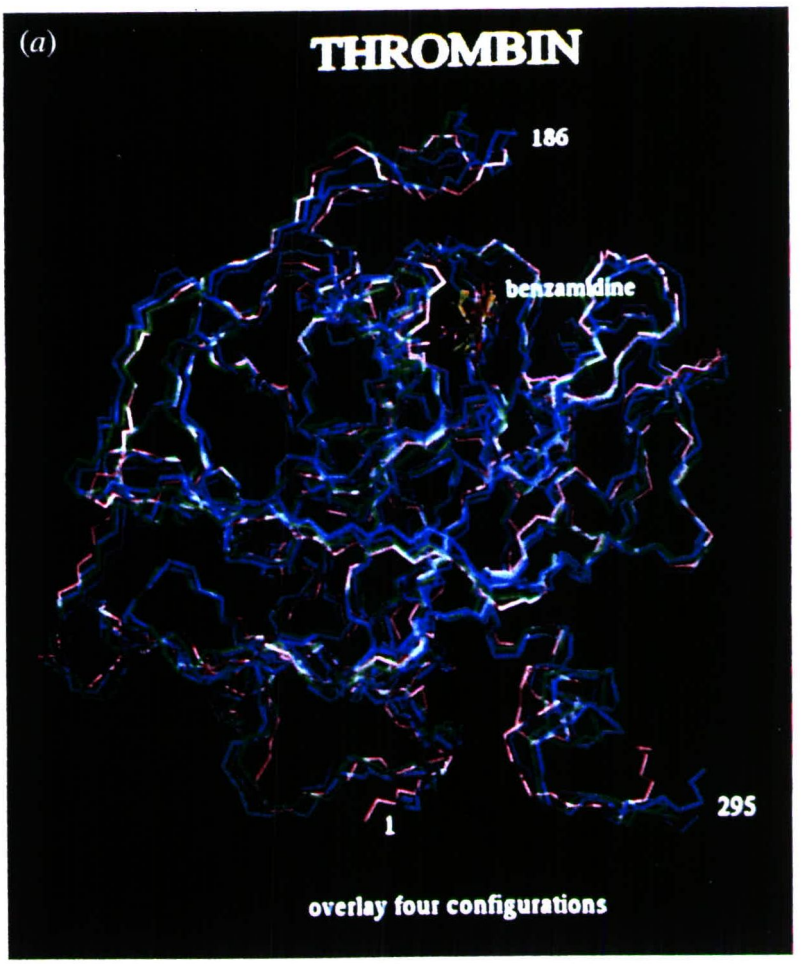
becomes a function of λ, and so does the free energy

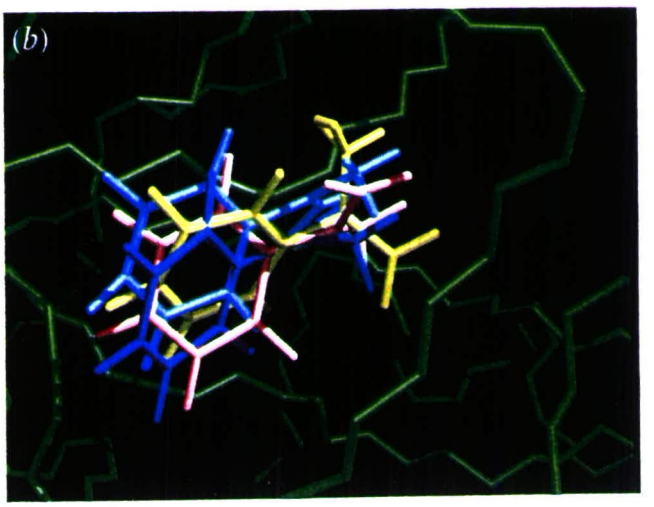
$$F(N,V,T;\lambda) = -k_B T \ln Z(N,V,T;\lambda). \tag{32}$$

The difference in free energy between the two states A and B becomes (for the same N, V, and T)

$$\Delta F_{BA} \equiv F(\lambda_B) - F(\lambda_A) = \int_{\lambda_A}^{\lambda_B} F'(\lambda)\,d\lambda \tag{33}$$

where $F'(\lambda) \equiv dF/d\lambda$. To simplify the notation, we drop the indication of constant





N, V and T for the remainder of the discussion. Differentiating (32) with respect to λ we find

$$F'(\lambda) = \iint \frac{\partial H(\boldsymbol{p},\boldsymbol{r};\lambda)}{\partial \lambda} P(\boldsymbol{p},\boldsymbol{r};\lambda)\,\mathrm{d}\boldsymbol{p}\,\mathrm{d}\boldsymbol{r}$$

$$= \left\langle \frac{\partial H(\lambda)}{\partial \lambda} \right\rangle_{\lambda} \tag{34}$$

where the probability of occurrence of molecular configuration (and momenta) $P(\boldsymbol{p},\boldsymbol{r};\lambda)$ and the ensemble average $\langle \ldots \rangle_{\lambda}$ defined in (28) have become dependent on λ. The advantage of formulae (33–34) over (25), (27) and (29) is clear: a relative free energy can be computed as an integral over the ensemble average of the derivative of the Hamiltonian with respect to the coupling parameter λ. This ensemble average does not suffer from the sampling problem which prohibits the computation of the absolute free energy via (25) or (27). Formulae (33–34) are called the thermodynamic integration formulae, which name originates in the analogy with obtaining free energy differences between states with different temperature or volume in thermodynamics.

The practical use of (33–34) involves three important issues.

1. The choice of the λ-dependence of $H(\boldsymbol{p},\boldsymbol{r};\lambda)$, which defines the pathway from state A to state B. Since the free energy is a state function, the change in free energy will in principle be independent of the path chosen. However, in practice this choice strongly influences the accuracy of the free energy difference obtained and the computational efficiency. The pathway should be chosen such that the relaxation time of the system with respect to the change in Hamiltonian and the time required to sample the ensemble are both minimized. This implies that the most direct path is not necessarily the most efficient (Mark *et al.* 1991). For a more elaborate discussion of the rather technical, but important considerations concerning the choice of $H(\boldsymbol{p},\boldsymbol{r};\lambda)$ we refer to the literature (van Gunsteren *et al.* 1993; Beutler *et al.* 1994; Mark & van Gunsteren, 1994*b*).
2. Determination of the ensemble averages $\langle \ldots \rangle_{\lambda}$ via computer simulation, in which proper equilibration and sampling are essential ingredients. At each λ-value the ensemble average $\langle \ldots \rangle_{\lambda}$ should be well converged and should include the molecular configurations that belong to the substate of the system which is defined by the particular value of λ. If λ changes, the range of these configurations may change and so will the entropy of the system. The range of configurations that can contribute to the ensemble average in a specific case is illustrated in Figures 4A and B. In practice, a primary if not the major source

---

Fig. 4. An illustration of the range of configurational states that make up the ensemble of accessible states that must be sampled to reliably estimate the derivative of free energy at a specific value λ. The figure shows an overlay of four randomly selected configurations from a 100ps simulation of a complex of benzamidine with human thrombin. Fluctuations of the protein backbone are shown in *A*. Motion of the bound ligand is shown in *B*.

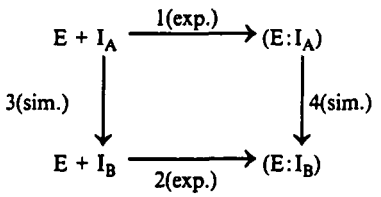
of error in free energy calculations is failure to sample a representative ensemble (Berendsen, 1991; Mark *et al.* 1994).

3. Evaluation of the integral over λ in (33). This is preferably done by computing the integrand (ensemble averages) at a few fixed λ-values and using simple numerical integration methods such as the trapezoidal rule, Simpson's rule or cubic spline integration (Mark *et al.* 1994).

4.2.2 *Thermodynamic cycles*

The next step when using the thermodynamic integration technique to calculate relative free energies, or binding constants of receptor-ligand complexes, or solubilities, is to formulate a so-called thermodynamic cycle (Beveridge & DiCapua, 1989). The basis on which the thermodynamic cycle approach rests is the fact that the free energy F is a thermodynamic state function. This means that as long as a system in equilibrium is changed in a reversible way, the change in free energy ΔF will be independent of the path of change. Therefore, along a closed path or cycle one has $\Delta F = 0$. This result implies that there are two possibilities of obtaining ΔF for a specific process. One may calculate it directly using the techniques discussed above along a path corresponding to the process, or one may design a cycle of which the specific process is only a part and calculate the ΔF of the remaining part of the cycle. The power of this thermodynamic cycle technique lies in the fact that on a computer also non-chemical processes such as the conversion of one type of atom into another type may be performed.

In order to visualize the method, we consider the relative binding of two inhibitors I_A and I_B to an enzyme E. The appropriate thermodynamic cycle for obtaining the relative binding constant is



(35)

where the symbol : means complex formation. The relative binding constant equals

$$K_2/K_1 = \exp[-(\Delta F_2 - \Delta F_1)/RT] \tag{36}$$

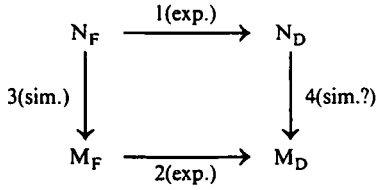
where R denotes the gas constant. However, simulation of processes *1* and *2* is virtually impossible, since it would involve the removal of many solvent molecules from the binding site of the inhibitor on the enzyme to be substituted by the inhibitor in a reversible manner. But, since (35) is a cycle we have

$$\Delta F_2 - \Delta F_1 = \Delta F_4 - \Delta F_3 \tag{37}$$

and, if the composition of inhibitor I_B is not too different from that of I_A, the desired result can be obtained by simulating the non-chemical processes *3* and *4* in a reversible manner.

Not every thermodynamic cycle one can think of can be used to obtain a relative

free energy difference by simulation using the method of thermodynamic integration (Shi *et al.* 1993). For example, the problem of protein stability can be described using the cycle (Dang *et al.* 1989; Tidor & Karplus, 1991)



(38)

where the folded (F) or denatured (D) form of a native (N) protein or a mutant (M) are the four systems involved. As before, the denaturation processes *1* and *2* cannot be carried out reversibly on a computer. If the mutation is not too large, process *3* may be carried out reversibly by simulation, but process *4* poses insurmountable problems: we do not know much about the denatured state of a protein, but the equilibration time and the sampling time of a partially unfolded protein certainly lie far beyond current computational limits. A similar situation is encountered when considering a thermodynamic cycle involving antibody-antigen binding. The equilibration and sampling of unbound linear peptides of sizeable length in aqueous solution still lies outside current computational possibilities.

4.2.3 *Use of restraints or constraints in a free energy calculation*

Since computer simulation of biomolecular complexes is still computationally expensive, one may try to reduce the extent of configuration space that is to be sampled by restraining the motion along part of the degrees of freedom of the system using a so-called restraining potential energy term, which is added to the Hamiltonian of the system. For example, spectroscopic data obtained by NMR measurements, such as Nuclear Overhauser Enhancement (NOE) intensities, J-coupling constants and chemical shifts may be used in such a restraining energy function term to generate an ensemble of protein structures that satisfies these experimental data (Kaptein *et al.* 1985; Torda *et al.* 1990, 1993; Harvey & van Gunsteren, 1993). In protein crystallography similar restraining terms involving measured structure factor amplitudes have been used for the same purpose (Brünger *et al.* 1987; Gros *et al.* 1990). Another example is the often used technique of restraining the motion of atoms that lie near the boundary of the system that is simulated in order to counteract the distortive forces due to the non-physical boundary and the vacuum beyond it. The atoms in the extended wall region are e.g. harmonically restrained to stationary positions (Brooks *et al.* 1985) or can be kept fixed (Berkowitz & McCammon, 1982). In the latter case one would rather speak of constraints than restraints. Constraints are also generally used to maintain fixed bond lengths or bond angles in a simulation to allow for larger integration time steps (van Gunsteren & Berendsen, 1990).

When restraints or constraints are used, the Hamiltonian of the system is different from the unrestrained or unconstrained one, which generally leads to a different ensemble being generated, and hence generally to different free energy

estimates. The introduction of constraints leads to metric tensor corrections (van Gunsteren *et al.* 1993). When applying bond length constraints, the correction is generally small and can be safely ignored. However, when a constrained bond length is changed as a function of λ in the coupling parameter approach, this change will contribute to the free energy change, irrespective of the method used to generate the ensemble or to maintain the constraint (Straatsma *et al.* 1992; van Gunsteren *et al.* 1993). The derivative of the (constrained) Hamiltonian with respect to the changing bond length must be determined and averaged over the ensemble. In other words, the force exerted by the environment to change the constraint, which is resisted by the constraint, must be determined and averaged. This force yields a contribution to the free energy: work done by the constraint force when changing the constraints. So, the contribution of a changing constraint (or restraint) to a free energy change can generally not be neglected. The amount of work done by the constraint (or restraint) forces will depend on the environment of the atoms involved in the changing constraint (or restraint). A growing bond length or bond angle may encounter a different resistance inside the protein than in solvent, so the amount of work done may be different. This implies that one generally cannot invoke cancellation of contributions from constraint (or restraint) forces to a free energy change computed along parallel legs of a thermodynamic cycle, e.g. processes 3 and 4 in (35), unless the environments of the changing constraint (or restraint) are very similar along both legs 3 and 4 of the cycle. The latter condition is in principle not fulfilled, since the basic idea of using a thermodynamic cycle is to compute the difference in free energy change of two different processes.

If the constraint or restraint term in the Hamiltonian does not depend on the coupling parameter λ, it will not yield a direct contribution to the free energy change, since its derivative with respect to λ equals zero. However, such λ-independent restraining may still strongly affect the free energy estimate obtained due to the restriction of the accessible configuration space of the system. For example, when applying position restraining to atoms in the extended wall region of the system, or when using instantaneous distance restraints based on NMR data, the motions of the atoms involved in these restraints are severely restricted, which restricts the ensemble of structures contributing to the energy and entropy. In such cases, the free energy estimate obtained is correct in terms of statistical mechanics, but not representative of the real molecular system involving the full atomic motions.

Summarizing we conclude that the application of constraints or restraints in simulations to obtain free energy estimates must be carefully considered both with respect to their direct contributions to free energy changes and with respect to their implications to a proper sampling of the configurations of the system that are relevant to the change in (free) energy and entropy that one aims to determine.

4.2.4 *Reliability and test of computed free energy differences*

When performing free energy calculations in practice, one would like to obtain an impression of the reliability of the obtained free energy differences. Below we list a number of possibilities to this end.

1. When computing ensemble averages $\langle \ldots \rangle_\lambda$, such as in (34) from a molecular simulation, the convergence of the ensemble average as a function of the number of simulation steps or time should be monitored at each chosen λ-value. If the simulation covers only a few picoseconds, only the fastest (bond, bond-angle) vibrations are sampled, and therefore reliably contribute to the free energy estimate, and other motions are neither equilibrated nor sufficiently sampled.
2. The addition of extra λ-values in the numerical integration over λ in (33) should not dramatically change the ΔF_{BA} value obtained so far. When creating or deleting atoms as a function of λ, accurate integration near the boundaries of the interval $[\lambda_A,\lambda_B]$ may require extra integration points (Mark *et al.* 1994).
3. When carrying out more than one change of a system, e.g. from inhibitor A to B and from inhibitor A to C, the quality of the equilibration, sampling and integration over λ can be tested by performing the change from inhibitor B to C, which closes a cycle:

$$\Delta F_{BA} + \Delta F_{CB} + \Delta F_{AC} = 0. \tag{39}$$

4. Repetition of individual simulations with different initial (equilibrium) configurations or velocities should yield the same result.
5. Small changes in the computational procedure should not affect the ΔF_{BA} value obtained (Shi *et al.* 1993).

We note that agreement with experimentally obtained free energy differences is a necessary, but not a sufficient condition for having obtained a reliable theoretical free energy estimate. In general, one is comparing a few numbers ($\Delta\Delta F$ values) calculated for a multi-dimensional system using many assumptions, approximations and parameters. One might well obtain good agreement between calculated and measured numbers for the wrong reasons, viz., accidental agreement, compensation of errors or adjustment of parameters (van Gunsteren, 1990).

Even in the case that entropic effects play no role in the differential binding of two ligands A and B to a receptor, use of the thermodynamic integration technique combined with a thermodynamic cycle will yield a more accurate estimate of the difference in binding energy than just taking the double difference of the average energies of the bound and unbound systems. In the latter case very large energies for slightly different systems are subtracted (twice) to obtain much smaller energy differences. The latter can be easily three orders of magnitude smaller than the system energies themselves (Fraternali & van Gunsteren, 1994), which leads to a dramatic loss of accuracy. Since the integrand (34) in the thermodynamic integration formulae (34–35) contains a derivative of the Hamiltonian with respect to the coupling parameter λ, only interactions that are chosen to be λ-dependent, will contibute to the (free) energy difference, which is, therefore, in general also orders of magnitude smaller than the system energy. Subtraction of large system energies can also be partly avoided by using the difference of the receptor-ligand energies for two different ligands as an estimate for the relative binding energy. However, such an estimate neglects the possible changes in intra-receptor and intra-ligand energies upon binding, apart from the

entropic contributions. So, even if entropic effects play no role in the binding, use of thermodynamic integration techniques is a most efficient way to obtain relative binding energy estimates for drug-receptor complexes.

Central to any rational design process is the ability to predict the effect of a proposed modification on the properties of interest. In the case of drug design the properties of interest depend directly on the associated changes in free energy. The basic methodology to predict changes in free energy based on molecular simulation techniques is well established and potentially highly accurate. Despite this, the quality of published free energy calculations varies greatly. Reliable free energy estimates from such calculations can only be obtained if the basic assumptions of equilibrium and proper sampling, on which the methods are based, are met.

4.2.5 *Free energy decomposition*

In the process of drug design it would be of great utility if a free energy of receptor-ligand binding could be decomposed in terms of contributions of particular (groups of) ligand atoms or specific types of interactions. Unfortunately, a meaningful separation of the free energy into specific components is, in general, not possible. The total free energy of a system can only be expressed in terms of a sum of components in so far as the total system can be separated into a set of independent subsystems. This is a direct consequence of the basic statistical mechanical definitions of free energy and entropy, (27–29) (van Gunsteren *et al.* 1993; Mark & van Gunsteren, 1994*a*).

A minimum requirement to express the total free energy as a sum of components is that the Hamiltonian of the system can be expressed as a sum of components, e.g.

$$H = H_1 + H_2 \tag{40}$$

where, for example, $H_1 = H_{bond}$ and $H_2 = H_{rest}$, or $H_1 = H_{residue}$ and $H_2 = H_{rest}$. Using (30), we see that the energy of the system can in principle be separated into components

$$E = \langle H \rangle = \langle H_1 + H_2 \rangle = \langle H_1 \rangle + \langle H_2 \rangle = E_1 + E_2. \tag{41}$$

The same is not true for either the free energy or the entropy. Combining (27) and (40) we obtain, apart from a constant term,

$$F = +k_B T \ln \langle e^{+U/k_B T} \rangle = +k_B T \ln \langle e^{+U_1/k_B T} e^{+U_2/k_B T} \rangle. \tag{42}$$

This expression can only be factorized to give

$$\begin{aligned} F &= +k_B T \ln \langle e^{+U_1/k_B T} \rangle \langle e^{+U_2/k_B T} \rangle \\ &= k_B T \ln \langle e^{+U_1/k_B T} \rangle + k_B T \ln \langle e^{+U_2/k_B T} \rangle \\ &= F_1 + F_2 \end{aligned} \tag{43}$$

if the factors $exp[+U_1/k_B T]$ and $exp[+U_2/k_B T]$ are not correlated, e.g. if U_1 and U_2 operate on different coordinates and are therefore not coupled. Thus, even if

the Hamiltonian of a system can be approximated by a linear combination of terms, it is generally not possible to associate each term with a given free energy.

If we expand both the exponential functions in (42) in terms of H_1 and H_2 and also use a Taylor expansion for the *ln* function, we find

$$F = \langle H_1 \rangle + \langle H_2 \rangle + (k_B T)^{-1} [\langle U_1 U_2 \rangle - \langle U_1 \rangle \langle U_2 \rangle] + O[(k_B T)^{-2}]$$
$$= E_1 + E_2 - TS. \tag{44}$$

The free energy can be expressed as a sum of energy components and a term containing all first and higher order correlations between H_1 and H_2, which represents the entropy of the system. The extent to which this term can be further separated is dependent on the degree to which the degrees of freedom of the system are correlated (Smith & van Gunsteren, 1994*b*).

In a number of recent studies an analysis of free energy components based on the thermodynamic integration technique has been presented (Kuczera *et al.* 1990; Tidor & Karplus, 1991; Simonson & Brunger, 1992; Miyamoto & Kollman, 1993; Prod'hom & Karplus, 1993), which is based on the formula

$$\Delta F_{BA} = \int_{\lambda_A}^{\lambda_B} \left\langle \frac{\partial H_1}{\partial \lambda} \right\rangle_\lambda \mathrm{d}\lambda + \int_{\lambda_A}^{\lambda_B} \left\langle \frac{\partial H_2}{\partial \lambda} \right\rangle_\lambda \mathrm{d}\lambda$$

$$= \int_{\lambda_A}^{\lambda_B} \frac{\iint \frac{\partial H_1(\boldsymbol{p}, \boldsymbol{r}; \lambda)}{\partial \lambda} \mathrm{e}^{-[H_1(\boldsymbol{p}, \boldsymbol{r}; \lambda) + H_2(\boldsymbol{p}, \boldsymbol{r}; \lambda)]/k_B T} \, \mathrm{d}\boldsymbol{p} \, \mathrm{d}\boldsymbol{r}}{\iint \mathrm{e}^{-[H_1(\boldsymbol{p}, \boldsymbol{r}; \lambda) + H_2(\boldsymbol{p}, \boldsymbol{r}; \lambda)]/k_B T} \, \mathrm{d}\boldsymbol{p} \, \mathrm{d}\boldsymbol{r}} \mathrm{d}\lambda$$

$$+ \int_{\lambda_A}^{\lambda_B} \frac{\iint \frac{\partial H_2(\boldsymbol{p}, \boldsymbol{r}; \lambda)}{\partial \lambda} \mathrm{e}^{-[H_1(\boldsymbol{p}, \boldsymbol{r}; \lambda) + H_2(\boldsymbol{p}, \boldsymbol{r}; \lambda)]/k_B T} \, \mathrm{d}\boldsymbol{p} \, \mathrm{d}\boldsymbol{r}}{\iint \mathrm{e}^{-[H_1(\boldsymbol{p}, \boldsymbol{r}; \lambda) + H_2(\boldsymbol{p}, \boldsymbol{r}; \lambda)]/k_B T} \, \mathrm{d}\boldsymbol{p} \, \mathrm{d}\boldsymbol{r}} \mathrm{d}\lambda$$

$$= \Delta F_{BA}(1) + \Delta F_{BA}(2). \tag{45}$$

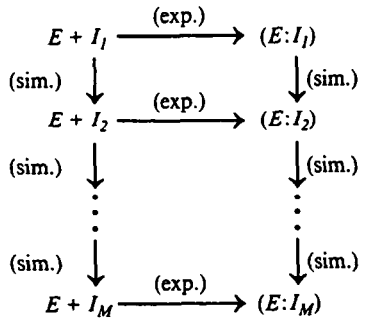
Formula (45) shows that the value of the components $\Delta F_{BA}(1)$ and $\Delta F_{BA}(2)$ will depend on the chosen λ-dependence of the Hamiltonian, which defines the pathway connecting system A with system B. This choice of pathway is not unique, and hence the free energy components calculated using (45) are not unique. A detailed analysis of free energy components based on (45) is, therefore, meaningless (Shi *et al.* 1993; Mark & van Gunsteren, 1994*a*).

4.2.6 *Free energy changes by extrapolation*

In terms of practical drug design the prediction of the relative free energy of binding of a single modified inhibitor will be of little use. It will, in almost all cases, be more efficient to synthesize and test the modified compound experimentally. The prediction of free energy trends for a large number of possible changes of the inhibitor would, in contrast, be of considerable help in guiding synthetic choices. This would be true as long as a range of modifications

could be treated in a single calculation and the calculation possesses reasonable predictive power. Such a method has been proposed recently (Gerber *et al.* 1993; Smith & van Gunsteren, 1994*a*).

If one is interested in the relative binding constant of a series of M different inhibitors $I_1, I_2, \ldots, I_M$ to an enzyme E, one would use $(M\text{-}1)$ thermodynamic cycles (35)



(46)

The $(M\text{-}1)$ relative free energy differences of binding can be obtained from simulating the $2(M\text{-}1)$ vertical processes in (46).

The free energy change in each vertical process can be obtained using the thermodynamic integration formula

$$F(I_{m+1}) - F(I_m) = \int_{\lambda_m}^{\lambda_{m+1}} F'(\lambda)\,\mathrm{d}\lambda \tag{47}$$

using (34) or higher-order derivatives of $F(\lambda)$ (Smith & van Gunsteren, 1994*a*). This is computationally costly, when the change of λ for each different process is separately carried out to compute (47). A considerable reduction of computational effort can be achieved by using for $F'(\lambda)$ in (47) an extrapolation from the point $\lambda = \lambda_m$ (van Gunsteren *et al.* 1993),

$$\begin{aligned} F'(\lambda = F'(\lambda_m) + F''(\lambda_m)(\lambda - \lambda_m)/1! + \ldots \\ + F^{(n)}(\lambda_m)(\lambda - \lambda_m)^{n-1}/(n-1)! + \ldots \end{aligned} \tag{48}$$

which yields

$$\begin{aligned} F(I_{m+1}) - F(I_m) &= F'(\lambda_m)[\lambda_{m+1} - \lambda_m] + F''(\lambda_m)[\lambda_{m+1} - \lambda_m]^2/2! + \ldots \\ &= \sum_{n=1}^{\infty} F^{(n)}(\lambda_m)[\lambda_{m+1} - \lambda_m]^n/n! \end{aligned} \tag{49}$$

Using (49) the free energy change in each vertical process in (46) is obtained using one λ-value, so by performing only one simulation to obtain $\langle \ldots \rangle_{\lambda_m}$. Of course, accuracy is lost using the approximation (49). However, the extrapolation formula (49) allows a further reduction of the computational effort in cases where the M inhibitors differ only slightly from each other: one inhibitor, say I_1, is selected to be simulated, and the free energy change to all other inhibitors I_m is computed using the extrapolation

$$F(I_m) - F(I_1) = \sum_{n=1}^{\infty} F^{(n)}(\lambda_1)[\lambda_m - \lambda_1]^n / n! \qquad (50)$$

for $m = 2,3,\ldots M$. Using (50) instead of (47) reduces the computational effort by about a factor MN_λ, where N_λ is the number of λ-values used in the numerical integration of (47).

This idea has been tested by Gerber *et al.* (1993) using only first derivatives $F'(\lambda_1)$ in (50). The use of the second and third derivatives $F''(\lambda_1)$ and $F'''(\lambda_1)$ in (50) improves the extrapolation approximation considerably in the case of dipolar changes in an aqueous environment (Smith & van Gunsteren, 1994*a*).

5. OUTLOOK

From a biophysical point of view drug design involves the optimisation or minimisation of the interaction between, in general, complex molecules in different molecular environments. The basic physical laws governing the behaviour of molecular systems are known: quantum and classical statistical mechanics. The formulae that express the free energy or entropy of a molecular system in terms of its Hamilton function or operator, or classically spoken its potential energy function, are available. The problem is, *only*, that the molecular systems of interest to drug design contain far too many degrees of freedom to handle exactly in terms of quantum or classical statistical mechanics. This leads to the necessity to make simplifying assumptions concerning the influence and importance of degrees of freedom, and to use approximations to the basic physical laws and to the nature of atomic interactions.

From a historical point of view drug design has evolved from establishing simple empirical relations between experimentally observed drug activity of specific molecules and particular physical or chemical properties of these molecules. With the advent of X-ray crystallography 3-dimensional structures of molecules became available which allowed for a correlation between drug activity and spatial molecular structure. The availability of molecular structures also offered the possibility to compute molecular energies, and so laid the bridge between fundamental physical calculations of (free) energies based on atomic and electronic degrees of freedom on the one hand, and molecular physical properties and the associated drug activity on the other hand.

The field of drug design on a molecular basis lies between two poles: (i) the detailed modelling of atomic and electronic degrees of freedom of molecular complexes based on fundamental physical laws, and (ii) the derivation of empirical models that relate a number of molecular properties to a large body of experimental observations using statistical methods to optimize the few model parameters.

Which are the forces driving the development of the methodology used in molecular drug design? Firstly, there is the still increasing number of molecular and protein 3-dimensional structures that become available each year from X-ray crystallography and multi-dimensional NMR spectroscopy. The databases of

molecular and protein structures contain a wealth of information that will be used more and more routinely in the drug design process when the appropriate software to extract and process the requested information becomes available and easy to use. Secondly, the rapid and continuing increase of computing power at constant price allows for the application of more and more complex and detailed molecular modelling techniques in the process of drug design. Last but not least, the development of drug design methodology is driven by demands of the drug market due to emerging new diseases and the decreased activity of established compounds against certain diseases.

These driving forces will partly determine the direction in which the methodology will develop and be applied. Database approaches will become more popular, and will be more widely used to derive empirical models based on statistical analysis of cause-effect relations. At the fundamental side the quantum mechanical treatments will cover larger molecules and include environmental effects in an average manner. Classical simulations based on semi-empirical force fields will be routinely used to study flexible molecular complexes and the associated entropic effects. The accuracy of such simulations will be enhanced by the steady improvement of the force fields used and the algorithms that are applied when searching and sampling conformational space. These improved force fields and molecular models will in turn be used in the empirical modelling of the relations between molecular properties and experimental data regarding drug activity.

Summarizing, we expect that both complex fundamental physical molecular modelling techniques and simple empirical models based on statistical analysis of experimental data concerning drug activity will continue to be developed. The drug designer will continue to have to choose between complex physically based methods which are relatively accurate, but too expensive or slow to be used in practice on the one hand, and simple empirical rule based models which are cheap and fast, but often not sufficiently accurate on the other hand. In other words she will continue to sail between the Scylla of using a too detailed and complex molecular model that does not give a timely answer at reasonable costs, and the Charybdis of neglecting too many degrees of freedom in a simple model that gives a quick and cheap answer which is useless due to its lack of accuracy.

6. REFERENCES

BERENDSEN, H. J. C. (1991). Incomplete equilibration: A source of error in free energy calculations. In *Proteins : Structure, Dynamics and Design* (ed. V. Renugopalakrishnan *et al.*), pp. 384–392. Leiden, NL: ESCOM Science Publishers B.V.

BERENDSEN, H. J. C. (1993). Electrostatic interactions. In *Computer Simulation of Biomolecular Systems : Theoretical and Experimental Applications*, vol. 2 (ed. W. F. van Gunsteren *et al.*), pp. 161–181. Leiden, NL: ESCOM Science Publishers B.V.

BERKOWITZ, M. & MCCAMMON, J. A. (1982). Molecular dynamics with stochastic boundary conditions. *Chem. Phys. Lett.* **90**, 215–217.

BEUTLER, T. C., MARK, A. E., VAN SCHAIK, R. C., GERBER, P. R. & VAN GUNSTEREN, W. F. (1994). Avoiding singularities and numerical instabilities in free energy calculations based on molecular simulations. *Chem. Phys. Letters* **222**, 529–539.

BEVERIDGE, D. L. & DI CAPUA, F. M. (1989). Free energy via molecular simulation: Applications to chemical and biochemical systems. *Ann. Rev. Biophys. Biophys. Chem.* **18**, 431–492.

BÖHM, H. J. (1992*a*). The computer program LUDI: A new method for the *de novo* design of enzyme inhibitors. *J. Comput.-Aided Mol. Design* **6**, 61–78.

BÖHM, H. J. (1992*b*). LUDI: rule-based automatic design of new substituents for enzyme inhibition leads. *J. Comput.-Aided Mol. Design* **6**, 593–606.

BOOBBYER, D. N. A., GOODFORD, P. J., MCWHINNIE, P. M. & WADE, R. C. (1989). New hydrogen-bond potentials for use in determining energetically favourable binding sites on molecules of known structure. *J. Med. Chem.* **32**, 1083–1094.

BOWEN-JENKINS, P. E., COOPER, D. L. & RICHARDS, W. G. (1985). *Ab initio* computation of molecular similarity. *J. Phys. Chem.* **89**, 2195–2197.

BROOKS, C. L., BRÜNGER, A. T. & KARPLUS, M. (1985). Active site dynamics in protein molecules: A stochastic boundary molecular-dynamics approach. *Biopolymers* **24**, 843–865.

BRÜNGER, A. T., KURIYAN, J. & KARPLUS, M. (1987). Crystallographic R factor refinement by molecular dynamics. *Science* **235**, 458–460.

BURT, C. & RICHARDS, W. G. (1990). Molecular similarity: the introduction of flexible fitting. *J. Comput.-Aided Mol. Design* **4**, 231–238.

BURT, C., RICHARDS, W. G. & HUXLEY, P. (1990). The application of molecular similarity calculations. *J. Comput. Chem.* **11**, 1139–1146.

CARBÓ, R. LEYDA, L. & ARNAU, M. (1980). How similar is one molecule to another? An electron density measure of similarity between two molecular structures. *Int. J. Quantum Chem.* **17**, 1185–1189.

CONNOLLY, M. L. (1983). Solvent-accessible surfaces of proteins and nucleic acids. *Science* **221**, 709–713.

COOPER, D. L. & ALLEN, N. L. (1989). A novel approach to molecular similarity. *J. Comput.-Aided Mol. Design* **3**, 253–259.

CRAMER, III, R. D., PATTERSON, D. E. & BUNCE, J. D. (1988). Comparative molecular field analysis (CoMFA). 1. Effect of shape on binding of steroids to carrier proteins. *J. Am. Chem. Soc.* **110**, 5959–5967.

CRIPPEN, G. M. & HAVEL, T. F. (1988). *Distance geometry and molecular conformation.* New York: Wiley.

DANG, L. X., MERZ, K. M. & KOLLMAN, P. A. (1989). Free energy calculations on protein stability: Thr-157 Val-157 mutation of T4 lysozyme. *J. Am. Chem. Soc.* **111**, 8505–8508.

DEAN, P. M. (1987). *Molecular Foundations of drug-receptor interaction.* Cambridge: Cambridge University Press.

DESJARLAIS, R. L., SHERIDAN, R. P., DIXON, J. S., KUNTZ, I. D. & VENKATARAGHAVAN, R. (1986). Docking flexible ligands to macromolecular receptors by molecular shape. *J. Med. Chem.* **29**, 2149–2153.

DE VLIEG, J., BERENDSEN, H. J. C. & VAN GUNSTEREN, W. F. (1989). An NMR based molecular dynamics simulation of the interaction of the *lac* repressor headpiece and its operator in aqueous solution. *Proteins* **6**, 104–127.

FOLKERS, G., MERZ, A. & ROGAN, D. (1993). CoMFA, scope and limitations. In *3D QSAR in Drug Design: Theory, Methods and Applications*, (ed. H. Kubinyi), pp. 583–618. Leiden NL: ESCOM Science Publishers B.V.

FRANKE, R. (1984). *Theoretical drug design methods.* Amsterdam: Elsevier.

FRATERNALI, F. & VAN GUNSTEREN, W. F. (1994). Conformational transitions of a

dipeptide in water: Effects of imposed pathways using umbrella sampling techniques. *Biopolymers* **34**, 347–355.

FRENKEL, D. (1993). Monte Carlo simulations: A primer. In *Computer Simulation of Biomolecular Systems: Theoretical and Experimental Applications*, vol. 2 (ed. W. F. van Gunsteren *et al.*), pp. 37–66. Leiden, NL: ESCOM Science Publishers B.V.

FRISCH, M. J., TRUCKS, G. W., HEAD-GORDON, M., GILL, P. M. W., WONG, M. W., FORESMAN, J. B., JOHNSON, B. G., SCHLEGEL, H. B., ROBB, M. A., REPLOGLE, E. S., GOMPERTS, R., ANDRES, J. L., RAGHAVACHARI, K., BINKLEY, J. S., GONZALEZ, C., MARTIN, R. L., FOX, D. J., DEFREES, D. J., BAKER, J., STEWART, J. J. P. & POPLE, J. A. (1992). Gaussian 92, Revision A. Pittsburgh: Gaussian Inc.

FUJITA, T., JUNKICHI, I. & HANSCH, C. (1964). A new substituent constant, π, derived from partition coefficients. *J. Am. Chem. Soc.* **86**, 5175–5180.

FUKUI, F., YANEZAWA, T. & SHINGU, H. (1952). A molecular orbital theory of reactivity in aromatic hydrocarbons. *J. Chem. Phys.* **20**, 722–725.

GELIN, B. R. (1993). Testing and comparison of empirical force fields: Techniques and problems. In *Computer Simulation of Biomolecular Systems: Theoretical and Experimental Applications*, vol. 2 (ed. W. F. van Gunsteren *et al.*), pp. 127–146. Leiden, NL: ESCOM Science Publishers B.V.

GERBER, P. R., MARK, A. E. & VAN GUNSTEREN, W. F. (1993). An approximate but efficient method to calculate free energy trends by computer simulation: Application to dihydrofolate reductase-inhibitor complexes. *J. Computer-Aided Molecular Design* **7**, 305–323.

GHOSE, A. K. & CRIPPEN G. M. (1985). Use of physico-chemical parameters in distance geometry and related three-dimensional quantitative structure-activity relationships: A demonstration using *Escherichia coli* dihydrofolate reductase inhibitors. *J. Med. Chem.* **28**, 3333–3346.

GILSON, M. K. & HONIG, B. (1991). The inclusion of electrostatic hydration energies in molecular mechanics calculations. *J. Comp.-Aided Mol. Des.* **5**, 5–20.

GOOD, A. C., HODGKIN, E. E. & RICHARDS, W. G. (1992*a*). Utilization of Gaussian functions for the rapid evaluation of molecular similarity. *J. Comp. Inf. Comput. Sci.* **32**, 188–191.

GOOD, A. C., HODGKIN, E. E. & RICHARDS, W. G. (1992*b*). Similarity screening of molecular data sets. *J. Comput.-Aided Mol. Des.* **6**, 513–520.

GOOD, A. C., SO, S.-S. & RICHARDS, W. G. (1993). Structure activity relationships from molecular similarity matrices. *J. Med. Chem.* **36**, 433–438.

GOODFORD, P. J. (1985). A computational procedure for determining energetically favourable binding sites on biologically important macromolecules. *J. Med. Chem.* **28**, 849–857.

GOODSELL, D. S. & OLSON, A. J. (1990). Automated docking of substrates to proteins by simulated annealing. *Proteins: Structure, Function and Genetics*, **8**, 195–202.

GROS, P., VAN GUNSTEREN, W. F. & HOL, W. G. J. (1990). Inclusion of thermal motion in crystallographic structures by restrained molecular dynamics. *Science* **249**, 1149–1152.

HAMMETT, L. P. (1940). *Physical Organic Chemistry*. New York: McGraw-Hill.

HANSCH, C., MUIR, R. M., FUJITA. T., MALONEY, P. P., GEIGER, F. & STREICH, M. (1963). The correlation of biological activity of plant growth regulators and Chloromycetin derivatives with Hammett constants and partition coefficients. *J. Am. Chem. Soc.* **85**, 2817–2824.

HANSCH, C. & FUJITA, T. (1964). ρ-σ-π Analysis. A method for the correlation of biological activity and chemical structure. *J. Am. Chem. Soc.* **86**, 1616–1626.

HARVEY, T. S. & VAN GUNSTEREN, W. F. (1993). The application of chemical shift calculation to protein structure determination by NMR, *Techniques in protein chemistry IV*. pp. 615–622. New York: Academic Press.

HEHRE, W. J., RADOM, L., VON RAGUÉ SCHLEYER, P. & POPLE, J. A. (1986). *Ab initio molecular orbital theory*. New York: Wiley.

HODGKIN, E. E. & RICHARDS, W. G. (1987). Molecular similarity based on electrostatic potential and electric field. *Int. J. Quantum Chem.: Quantum Biol. Symp.* **14**, 105–110.

HODGKIN, E. E., MILLER, A. & WHITTAKER, M. (1993). A Monte Carlo pharmacophore generation procedure: Application to the human PAF receptor. *J. Comp.-Aided Mol. Design* **7**, 515–534.

HOPFINGER, A. J. & BURKE, B. J. (1990). Molecular shape analysis: A formalism to quantitatively establish spatial molecular similarity. In *Concepts and applications of molecular similarity* (ed. M. A. Johnson *et al.*), pp. 173–209. New York: Wiley.

HUBER, T., TORDA, A. E. & VAN GUNSTEREN, W. F. (1994). Local elevation: A method for improving the searching properties of molecular dynamics simulation. *J. Comp.-Aided Mol. Design*, in press.

KANG, Y. K., GIBSON, K. D., NÉMETHY, G. & SCHERAGA, H. A. (1988). Free energies of hydration of solute molecules. 4. Revised treatment of the hydration shell model. *J. Phys. Chem.* **92**, 4739–4742.

KAPTEIN, R., ZUIDERWEG, E. R. P., SCHEEK, R. M., BOELENS, R. & VAN GUNSTEREN, W. F. (1985). A protein structure from nuclear magnetic resonance data: *lac* repressor headpiece. *J. Mol. Biol.* **182**, 179–182.

KELLOGG, G. E., SEMUS, S. F. & ABRAHAM, D. J. (1991). HINT: A new method of empirical hydrophobic field calculation for CoMFA. *J. Comp. Aided. Mol. Design* **5**, 545–552.

KIM, K. H. (1991). A novel method of describing hydrophobic effects directly from 3D structures in 3D-quantitative structure-activity relationship studies. *Med. Chem. Res.* **1**, 259–264.

KIM, H. K., GRECO, G., NOVELLINO, E., SILIPO, C. & VITTORIA, A. (1993). Use of the hydrogen bond potential function in a comparative molecular field analysis (CoMFA) on a set of benzodiazepines. *J. Comp. Aided. Mol. Design* **7**, 263–280.

KING, P. M. (1993). Free energy via molecular simulation: A primer. In *Computer Simulation of Biomolecular Systems: Theoretical and Experimental Applications*, vol. 2 (ed. W. F. van Gunsteren *et al.*), pp. 267–314. Leiden, NL: ESCOM Science Publishers B.V.

KING, P. M., SPYCHER, R. M. & VAN GUNSTEREN, W. F. (1993). Structure elucidation from rotation spectra: a penalty function approach. *Chem. Phys. Letters* **203**, 88–92.

KLEBE, G. & ABRAHAM, U. (1993). On the prediction of binding properties of drug molecules by comparative molecular field analysis. *J. Med. Chem.* **36**, 70–80.

KUBINYI, H. (1993). *QSAR: Hansch analysis and related approaches*. Weinheim, Germany: VCH.

KUCZERA, K., GAO, J., TIDOR, B. & KARPLUS, M. (1990). Free energy of sickling: A simulation analysis. *Proc. Natl. Acad. Sci.* **87**, 8481–8485.

KUNTZ, I. D. (1992). Structure-based strategies for drug design and discovery. *Science* **257**, 1078–1082.

KUNTZ, I. D., BLANEY, J. M., OATLEY, S. J., LANGRIDGE, R. & FERRIN, T. E. (1982). A

geometric approach to macromolecular-ligand interactions. *J. Mol. Biol.* **161**, 269–288.

LEACH, A. R., PROUT, K. & DOLATA, D. P. (1990). The application of artificial intelligence to the conformational analysis of strained molecules. *J. Comp. Chem.* **11**, 680–693.

MARK, A. E. & VAN GUNSTEREN, W. F. (1994*a*). Decomposition of the free energy of a system in terms of specific interactions: Implications for theoretical and experimental studies. *J. Mol. Biol.* **240**, 167–176.

MARK, A. E. & VAN GUNSTEREN, W. F. (1994*b*). Free energy calculations in drug design: A practical guide. To appear in *Perspectives in Drug Design*. Proceedings of the 9th Intl. Roundtable at Turnberry, Scotland.

MARK, A. E., VAN GUNSTEREN, W. F. & BERENDSEN, H. J. C. (1991). Calculation of relative free energy via indirect pathways. *J. Chem. Phys.* **94**, 3808–3816.

MARK, A. E., VAN HELDEN, S. P., SMITH, P. E., JANSSEN, L. H. M. & VAN GUNSTEREN, W. F. (1994). Convergence properties of free energy calculations: α-cyclodextrin complexes as a case study. *J. Am. Chem. Soc.*, **116**, 6293–6302.

MARTIN, Y. C. (1978). *Quantitative drug design: A critical introduction*. New York: Marcel Dekker.

MARTIN, Y. C. (1991). Computer-assisted rational drug design. *Methods in Enzymology* **203**, 587–613.

MATTOS, C., RASMUSSEN, B., DING, X., PETSKO, G. A. & RINGE, D. (1994). Analogous inhibitors of elastase do not always bind analogously. *Nature: Struct. Biol.* **1**, 55–58.

MIYAMOTO, S. & KOLLMAN, P. A. (1993). Absolute and relative binding free energy calculations of the interaction of biotin and its analogs with streptavidin using molecular dynamics/free energy perturbation approaches. *Proteins: Struct. Funct. Genet.* **16**, 226–245.

MOON, J. B. & HOWE, J. W. (1991). Computer design of bioactive molecules: A method for receptor based de novo ligand design. *Proteins: Struct. Funct. Genet.* **11**, 314–328.

NISHIBATA, Y. & ITAI, A. (1991). Automatic creation of drug candidate structures based on receptor structure. Starting point for artificial lead generation. *Tetrahedron* **47**, 8985–8990.

OOI, W., OOBATAKE, M. NÉMETHY, G. & SCHERAGA, H. A. (1987). Accessible surface areas as a measure of the thermodynamic parameters of hydration of peptides. *Proc. Natl. Acad. Sci. USA* **84**, 3086–3090.

PARR, R. G. & YANG, W. (1989). *Density-Functional Theory of Atoms and Molecules*. Oxford: Oxford University Press.

PEARLMAN, D. A. & MURCKO, M. A. (1993). Concepts: New dynamic algorithm for *de novo* drug suggestion. *J. Comp. Chem.* **14**, 1184–1193.

PICKERSGILL, R. W. (1988). A rapid method of calculating charge-charge interaction energies in proteins. *Protein Engineering* **2**, 247–248.

PROD'HOM, B. & KARPLUS, M. (1993). The nature of the ion binding interactions in EF-hand peptide analogs: free energy simulation of Asp to Asn mutations. *Protein Engineering* **6**, 585–592.

RICHARDS, W. G. (1983). *Quantum Pharmacology* 2nd edn. London: Butterworths.

ROTSTEIN, S. H. & MURCKO, M. A. (1993). GenStar: A method for *de novo* drug design. *J. Comp. Aided. Mol. Design* **7**, 23–43.

RUSINKO III, A., SKELL, J. M., BALDUCCI, R., MCGARITY, C. M. & PEARLMAN, R. S. (1988). CONCORD, a program for the rapid generation of high quality approximate

3-dimensional molecular structures. The University of Texas and Austin and Tripos Associates, St. Louis, MO, USA.

RUTENBER, E., FAUMAN, E. B., KEENAM, R. J., FONG, S., FURTH, P. S., ORTIZ DE MONTELLANO, P. R., MENG, E., KUNTZ, I. D., DE CAMP, D. L., SALTO, R., ROSE, J. R., CRAIK, C. S. & STROUD, R. M. (1993). Structure of a non-peptide inhibitor complexed with HIV-1 protease. *J. Biol. Chem.* **268**, 15343–15346.

SCHEEK, R. M., TORDA, A. E., KEMMINK, J. & VAN GUNSTEREN, W. F. (1991). Structure determination by NMR: The modelling of NMR parameters as ensemble averages. In *Computational Aspects of the Study of Biological Macromolecules by Nuclear Magnetic Resonance Spectroscopy* (ed. J. C. Hoch *et al.*), pp. 209–217. New York: NATO ASI Series **A225**, Plenum Press.

SCHERAGA, H. A. (1993). Searching conformational space. In *Computer Simulation of Biomolecular Systems: Theoretical and Experimental Applications*, vol. 2 (ed. W. F. van Gunsteren *et al.*), pp. 231–248. Leiden, NL: ESCOM Science Publishers B.V.

SCHIFFER, C. A., CALDWELL, J. W., STROUD, R. M. & KOLLMAN, P. A. (1992). Inclusion of solvation free energy with molecular mechanics energy: alanyl dipeptide as a test case. *Protein Science* **1**, 396–400.

SHARP, K. A. (1991). Incorporating solvent and ion screening into molecular dynamics using the finite-difference Poisson-Boltzmann method. *J. Comput. Chem.* **12**, 454–468.

SHARP, K. A. (1993). Inclusion of solvent effects in molecular mechanics force fields. In *Computer Simulation of Biomolecular Systems: Theoretical and Experimental Applications*, vol. 2 (ed. W. F. van Gunsteren *et al.*), pp. 147–160. Leiden, NL: ESCOM Science Publishers B.V.

SHERIDAN, R. P., NILAKANTAN, R., DIXON, J. S. & VENKATARAGHAVEN, R. (1986). The ensemble approach to distance geometry: Application to the nicotinic pharmacophore. *J. Med Chem.* **29**, 899–906.

SHI, Y. Y., MARK, A. E., WANG, C. X., HUANG, F., BERENDSEN, H. J. C. & VAN GUNSTEREN, W. F. (1993). Can the stability of protein mutants be predicted by free energy calculations? *Protein Engineering* **6**, 289–295.

SHOICHET, B. K., BODIAN, D. L. & KUNTZ, I. D. (1992). Molecular docking using shape descriptors. *J. Comput. Chem.* **13**, 380–397.

SHOICHET, B. K., STROUD, R. M., SANTI, D. V., KUNTZ, I. D. & PERRY, K. M. (1993). Structure-based discovery of inhibitors of thymidylate synthase. *Science* **259**, 1445–1450.

SILVERMAN, R. B. (1992). *The organic chemistry of drug design and drug action.* San Diego, USA: Academic Press.

SIMONSON, T. & BRUNGER, A. T. (1992). Thermodynamics of protein-peptide interactions in the ribonuclease-S system studied by molecular dynamics and free energy calculations. *Biochemistry* **31**, 8661–8674.

SMITH, P. E. & VAN GUNSTEREN, W. F. (1993). Methods for the evaluation of long-range electrostatic forces in computer simulations of molecular systems. In *Computer Simulation of Biomolecular Systems: Theoretical and Experimental Applications*, vol. 2 (ed. W. F. van Gunsteren *et al.*), pp. 182–212. Leiden, NL: ESCOM Science Publishers B.V.

SMITH, P. E. & VAN GUNSTEREN, W. F. (1994*a*). Predictions of free energy differences from a single simulation of the initial state. *J. Chem. Phys.* **100**, 577–585.

SMITH, P. E. & VAN GUNSTEREN, W. F. (1994*b*). When are free energy components meaningful? *J. Phys. Chem.* (in press).

SOLMAJER, T. & MEHLER, E. L. (1991). Electrostatic screening in molecular dynamics simulations. *Protein Engineering* **4**, 911–917.

STEWART, J. J. P. (1990). Mopac: A semiempirical molecular orbital program. *J. Comput.-Aided Mol. Design* **4**, 1–105.

STILL, W. C., TEMPCZYK, A., HAWLEY, R. C. & HENDRICKSON, T. (1990). Semianalytical treatment of solvation for molecular mechanics and dynamics. *J. Am. Chem. Soc.* **112**, 6127–6129.

STODDARD, B. L. & KOSHLAND JR, D. E. (1992). Prediction of the structure of a receptor-protein complex using a binary docking method. *Nature* **358**, 774–776.

STOUTEN, P. F. W., FRÖMMEL, C., NAKAMURA, H. & SANDER, C. (1993). An effective solvation term based on atomic occupancies for use in protein simulations. *Molecular Simulation* **10**, 97–120.

STRAATSMA, T. P., ZACHARIAS, M. & MCCAMMON, J. A. (1992). Holonomic constraint contributions to free energy differences from thermodynamic integration molecular dynamics simulations. *Chem. Phys. Lett.* **196**, 297–302.

STRAATSMA, T. P., ZACHARIAS, M. & MCCAMMON, J. A. (1993). Free energy difference calculations in biomolecular systems. In *Computer Simulation of Biomolecular Systems: Theoretical and Experimental Applications*, vol. 2 (ed. W. F. van Gunsteren *et al.*), pp. 349–370. Leiden, NL: ESCOM Science Publishers B.V.

SUZUKI, T. & KUDO, Y. (1990). Automatic log P estimation based on combined additive modelling methods. *J. Comp. Aid. Molec. Design* **4**, 155–198.

TAFT JR, R. W. (1956). In *Steric effects in organic chemistry* (ed. M. S. Newman), pp. 556. New York: Wiley.

TANFORD, C. (1973). *The hydrophobic effect: Formation of micelles and biological membranes*. New York: Wiley.

TIDOR, B. & KARPLUS, M. (1991). Simulation analysis of the stability mutant R96H of T4 lysozyme. *Biochemistry* **30**, 3217–3228.

TORDA, A. E., BRUNNE, R. M., HUBER, T., KESSLER, H. & VAN GUNSTEREN, W. F. (1993). Structure refinement using time-averaged J-coupling constant restraints. *J. Biomol. NMR* **3**, 55–66.

TORDA, A. E., SCHEEK, R. M. & VAN GUNSTEREN, W. F. (1989). Time-dependent distance restraints in molecular dynamics simulations. *Chem. Phys. Letters* **157**, 289–294.

TORDA, A. E., SCHEEK, R. M. & VAN GUNSTEREN, W. F. (1990). Time-averaged Nuclear Overhauser Effect distance restraints applied to tendamistat. *J. Mol. Biol.* **214**, 223–235.

VAN GUNSTEREN, W. F. (1990). On testing theoretical models by comparison of calculated with experimental data. In *Studies in Physical and Theoretical Chemistry*, vol. 71 (ed. J.-L. Rivail), pp. 463–478. Amsterdam: Elsevier.

VAN GUNSTEREN, W. F. (1993). Molecular dynamics and stochastic dynamics simulation: A primer. In *Computer Simulation of Biomolecular Systems: Theoretical and Experimental Applications*, vol. 2 (ed. W. F. van Gunsteren *et al.*), pp. 3–36. Leiden, NL: ESCOM Science Publishers B.V.

VAN GUNSTEREN, W. F. & BERENDSEN, H. J. C. (1990). Computer simulation of molecular dynamics: Methodology, applications and perspectives in chemistry. *Angew. Chem. Int. Ed. Engl.* **29**, 992–1023.

VAN GUNSTEREN, W. F. & MARK, A. E. (1992*a*). On the interpretation of biochemical data by molecular dynamics computer simulation. *Eur. J. Biochem.* **204**, 947–961.

VAN GUNSTEREN, W. F. & MARK, A. E. (1992*b*). Prediction of the activity and stability effects of site-directed mutagenesis on a protein core. *J. Mol. Biol.* **227**, 389–395.

VAN GUNSTEREN, W. F., BEUTLER, T. C., FRATERNALI, F., KING, P. M., MARK, A. E. & SMITH, P. E. (1993). Computation of free energy in practice: choice of approximations and accuracy limiting factors. In *Computer simulation of biomolecular systems: theoretical and experimental applications*, Vol 2 (ed. W. F. van Gunsteren *et al.*), pp. 315–348. Leiden, NL: ESCOM Science Publishers B.V.

VAN GUNSTEREN, W. F., BRUNNE, R. M., GROS, P., VAN SCHAIK, R. C., SCHIFFER, C. A. & TORDA, A. E. (1994*a*). Accounting for molecular mobility in structure determination based on NMR spectroscopic and X-ray diffraction data. *Methods in Enzymology* **239**, 619–654.

VAN GUNSTEREN, W. F., LUQUE, F. J., TIMMS, D. & TORDA, A. E. (1994*b*). Molecular mechanics in biology: From structure to function: Taking account of solvation. *Ann. Rev. Biophys. Biomol. Structure* **23**, 847–863.

VAN SCHAIK, R. C., BERENDSEN, H. J. C., TORDA, A. E. & VAN GUNSTEREN, W. F. (1993). A structure refinement method based on molecular dynamics in four spatial dimensions. *J. Mol. Biol.* **234**, 751–762.

VAN SCHAIK, R. C., VAN GUNSTEREN, W. F. & BERENDSEN, H. J. C. (1992). Conformational search by potential energy annealing: Algorithm and application to cyclosporin A. *J. of Comp.-Aided Mol. Design* **6**, 97–112.

VERLOOP, A., HOOGENSTRAATEN, W. & TIPKER, J. (1976). Development and application of new steric substituent parameters in drug design. In *Drug Design*, vol III (ed. E. J. Ariens), New York: Academic Press.

VILA, J., WILLIAMS, R. L., VÀSQUEZ, M. & SCHERAGA, H. A. (1991). Empirical solvation models can be used to differentiate native from near-native conformations of bovine pancreatic trypsin inhibitor. *Proteins* **10**, 199–218.

WADE, R. C., CLARK, K. J. & GOODFORD, P. J. (1993). Further development of hydrogen bond functions for use in determining energetically favourable binding sites on molecules of known structure. 1. Ligand probe groups with the ability to form two hydrogen bonds. *J. Med. Chem.* **36**, 140–147.

WADE, R. C. & GOODFORD, P. J. (1993). Further development of hydrogen bond functions for use in determining energetically favourable binding sites on molecules of known structure. 2. Ligand probe groups with the ability to from more than two hydrogen bonds. *J. Med. Chem.* **36**, 148–158.

WALLER, C. L. & MARSHALL, G. (1993). Three-dimensional quantitative structure-activity relationship of angiotensin-converting enzyme and thermolysin inhibitors. II. A comparison of CoMFA models incorporating molecular orbital fields and desolvation free energies based on active-analog and complementary-receptor-field alignment rules. *J. Med. Chem.* **36**, 2390–2403.

WESSON, L. & EISENBERG, D. (1992). Atomic solvation parameters applied to molecular dynamics of proteins in solution. *Protein Science* **1**, 227–235.

WOLD, S., RUHE, A., WOLD, H. & DUNN, III, W. J. (1984). The collinearity problem in linear regression. The partial least squares (PLS) approach to generalized inverses. *SIAM J. Sci. Stat. Comput.* **5**, 735–743.